DOCUMENT RESUME

ED 477 359 TM 034 954

AUTHOR	Rudner, Lawrence M., Ed.; Schaefer, William D., Ed.
TITLE	Practical Assessment, Research and Evaluation, 2002-2003.
INSTITUTION	ERIC Clearinghouse on Assessment and Evaluation, College Park, MD.; Maryland Univ., College Park. Dept. of Measurement, Statistics & Evaluation.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
ISSN	ISSN-1531-7714
PUB DATE	2000-12-00
NOTE	80p.; "Practical Assessment, Research & Evaluation" is an electronic-only journal covered on an article-by-article basis in "Current Index to Journals in Education" (CIJE). For these 10 articles, see TM 525 172-181.
AVAILABLE FROM	For full text: http://ericae.net/pare.
PUB TYPE	Collected Works - Serials (022) -- ERIC Publications (071)
JOURNAL CIT	Practical Assessment, Research and Evaluation; n1-10 2002-2003
EDRS PRICE	EDRS Price MF01/PC04 Plus Postage.
DESCRIPTORS	Bilingual Education; Bilingual Students; Cheating; *Educational Research; Elementary Secondary Education; *Evaluation Methods; Formative Evaluation; Limited English Speaking; *Online Systems; Regression (Statistics); *Research Methodology; Sex Differences; Surveys

ABSTRACT

This document consists of the first 10 articles of volume 8 of the electronic journal "Practical Assessment, Research & Evaluation" published in 2002-2003: (1) "Using Electronic Surveys: Advice from Survey Professionals" (David M. Shannon, Todd E. Johnson, Shelby Searcy, and Alan Lott); (2) "Four Assumptions of Multiple Regression That Researchers Should Always Test" (Jason W. Osbourne and Elaine Waters); (3) "Analyzing Online Discussions: Ethics, Data, and Interpretation" (Sarah K. Brem); (4) "Language Ability Assessment of Spanish-English Bilinguals Future Directions" (Ellen Stubbe Kester and Elizabeth D. Pena); (5) "Male and Female Differences in Self-report Cheating" (James A. Athanasou and Olabisi Olasehinde); (6) "Notes on the Use of Data Transformations" (Jason W. Osbourne); (7) "Linguistic Simplification: A Promising Test Accommodation for LEP Students?" (Charles W. Stansfield); (8) "Evaluating Classroom Communication: In Support of Emergent and Authentic Frameworks in Second Language Assessment" (Miguel Mantero); (9) "The Concept of Formative Assessment" (Carol Boston); and (10) "Vertical Equating for State Assessments: Issues and Solutions in Determination of Adequate Yearly Progress and School Accountability" (Robert W. Lissitz and Huynh Huynh). (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Volume 8, 2000

ISSN 1531-7714

ED 477 359

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal

(Articles 1-10)

TM034954

Practical Assessment, Research and Evaluation
is listed among the ejournals in education at the website for the
AERA SIG "Communications Among Researchers"

BEST COPY AVAILABLE

Board of Editors

Kathryn Alvestad, Calvert County (MD) Public Schools

Sarah Brem
Arizona State Univ

Filip Dochy,
Univ of Maastricht

Kurt F. Geisinger,
St. Thomas University

Gene V Glass,
Arizona State Univ

Arlen R. Gullickson
Western Michigan University

Robin K. Henson,
University of North Texas

Robert Marzano, Mid-continent Research for Education and Learning

M. Kevin Matter
Cherry Creek (C) Schools

Donna.Mertens
Gallaudet University

Denise McKeon,
National Education Assoc

Mark Moody,
Consultant

Joe O'Reilly,
Mesa (AZ) Public Schools

Albert Oosterhof
Florida State University

Jason W. Osborne,
North Carolina State Univ

Dennis Roberts,
Penn State Univ

Michael Russell,
Boston College

Bruce Thompson,
Texas A&M

Marie Miller-Whitehead
Consultant

Lynn Winters, Long Beach Unified Public Schools Editorial Board

Lawrence M. Rudner,
Univ of Maryland, co-Editor

William D. Schafer,
Univ of Maryland, co-Editor

Steven L. Wise,
James Madison University

Practical Assessment, Research and Evaluation (PARE) is an on-line journal published by the **ERIC Clearinghouse on Assessment and Evaluation (ERIC/AE) and the Department of Measurement, Statistics, and Evaluation at the University of Maryland, College Park**. Its purpose is to provide education professionals access to refereed articles that can have a positive impact on assessment, research, evaluation, and teaching practice, especially at the local education agency (LEA) level.

Manuscripts published in *Practical Assessment, Research and Evaluation* are scholarly syntheses of research and ideas about issues and practices in education. They are designed to help members of the community keep up-to-date with effective methods, trends and research developments. While they are most often prepared for practitioners, such as teachers, administrators, and assessment personnel who work in schools and school systems, *PARE* articles can target other audiences, including researchers, policy makers, parents, and students.

Manuscripts to be considered for *Practical Assessment, Research and Evaluation* should be short, 2000-8000 words or about eight pages in length, exclusive of tables and references. They should conform to the stylistic conventions of the American Psychological Association (APA). See the Policies section of this web site for technical specifications and a list of suggested topics. Manuscripts should be submitted electronically to pare2@ericae.net. Articles appearing in *Practical Assessment, Research and Evaluation* also become available in the ERIC database through the ERIC Digest Series. Many articles published in *PARE* were previously published as part of the *ERIC/AE Digest Series.*

Permission is granted to distribute any article in this journal for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used.

Practical Assessment, Research and Evaluation is listed among the ejournals in education by the Scholarly Publishing and Academic Resources Coalition(SPARC), and the website for the AERA SIG "Communications Among Researchers".

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal

Articles Pages

1. Using Electronic Surveys: Advice from Survey Professionals
Shannon, David M., Johnson, Todd E., Searcy, Shelby & Alan Lott 11

2. Four Assumptions of Multiple Regression That Researchers Should Always Test
Osborne, Jason and Elaine Waters 7

3. Analyzing Online Discussions: Ethics, Data, and Interpretation
Brem, Sarah 8

4. Language Ability Assessment of Spanish-english Bilinguals: Future Directions
Kester, Ellen Stubbe and Elizabeth D. Peña 6

5. Male and Female Differences in Self-report Cheating
Athanasou, James A. and Olabisi Olasehinde 14

6. Notes on the use of data transformations
Osborne, Jason 9

7. Linguistic Simplification: A Promising Test Accommodation for LEP Students?
Stansfield, Charles W. 4

8. Evaluating Classroom Communication: In Support of Emergent and Authentic Frameworks in Second Language Assessment
Mantero, Miguel 6

9. The Concept of Formative Assessment
Boston, Carol 5

10. Vertical Equating for State Assessments: Issues and Solutions in Determination of Adequate Yearly Progress and School Accountability
Lissitz, Robert W. & Huynh Huynh 6

Home | Articles | Subscribe | Review | Policies

Volume: 8 7 6 5 4 3 2 1

Practical Assessment, Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ☐ Go

Copyright 2002, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Shannon, David M., Johnson,Todd E., Searcy, Shelby , Alan Lott (2002). Using electronic surveys: advice from survey professionals. *Practical Assessment, Research & Evaluation*, 8(1). Available online: http://ericae.net/pare/getvn.asp?v=8&n=1. This paper has been viewed 6200 times since 1/2/02.

- Find similar papers in
  ERICAE Full Text Library
  Pract Assess, Res & Eval
  ERIC RIE & CIJE 1990-
  ERIC On-Demand Docs
- Find articles in ERIC written by
  Shannon, David M.
  Johnson,Todd E.
  Searcy, Shelby
  Alan Lott

Using Electronic Surveys: Advice from Survey Professionals

David M. Shannon, Auburn University
Todd E. Johnson, Auburn University
Shelby Searcy, Huntington College
Alan Lott, Auburn University

Abstract

The study reports the perceptions and recommendations of sixty-two experienced survey researchers from the American Educational Research Association regarding the use of electronic surveys. The most positive aspects cited for the use of electronic surveys were reduction of costs (i.e., postage, phone charges), the use of electronic mail for pre-notification or follow-up purposes, and the compatibility of data with existing software programs. These professionals expressed limitations in using electronic surveys pertaining to the limited sampling frame as well as issues of confidentiality, privacy, and the credibility of the sample. They advised that electronic surveys designed with the varied technological background and capabilities of the respondent in mind, follow sound principles of survey construction, and be administered to pre-notified, targeted populations with published email addresses.

There has been an extensive amount of research focused on principles of survey design and factors influencing response to mail and telephone surveys (Babbie, 1990; Baruch, 1999; Dillman, 1978; Herberlein & Baumgartner, 1978; Fowler, 1993; Lavrakas, 1993; Linsky, 1975; Sudman and Bradburn, 1982;Yu & Cooper, 1983). From the efforts of survey researchers, we have discovered important considerations when designing survey instruments including the importance of the first question, grouping and sequencing of questions, establishing a respondent-pleasing vertical flow of items in the survey, and having clear specific directions. We have also learned the

BEST COPY AVAILABLE

importance of implementation components like pre-notification of respondents, personalized cover letters, incentives, return postage, and multiple contacts to reach respondents and generate higher response rates.

The Internet has greatly impacted the field of survey research as the number of electronically-administered surveys continue to grow. Unlike traditional mail and telephone surveys, it's not certain what principles should guide the construction and implementation of electronic surveys. Preliminary efforts have suggested many of the same principles apply to electronic surveys (Cook, Heath & Thompson, 2000; Dillman, 2000; Dillman & Bowker, 2000; Dillman, Tortora, & Bowker, 1998; Schaeffer & Dillman, 1998; Shannon & Bradshaw, in press). Additional research is needed to refine these principles and use them most effectively with the design and implementation of electronic surveys, especially given the wide variety of formats used to conduct electronic surveys. We will discuss three common forms of electronic surveys below.

Electronic surveys have taken on a variety of forms from simple email surveys to sophisticated web survey systems. Early forms of electronic surveys existed in the form of the disk-by-mail format (Couper & Nichols, 1998). Using this approach, a disk that contained the survey is mailed to respondents, who are instructed to open the file, complete the survey, and mail the disk back to the researcher. Bowers (1999) describes these surveys as having the capability of guiding the respondent interactively through the survey and including very complex skip patterns or rotation logic. This approach can offer many innovative features beyond traditional mail and telephone surveys, but it does require costs and time in terms of programming and distribution of the survey. However, this approach is restricted by the technological capacity of the respondent's computer. In addition, Bowers (1999) warns that respondents may be reluctant to download files in fear that they may contain viruses.

A second type of electronic survey is the e-mail survey. These surveys are typically contained within an e-mail message or as an attached file (Bradley, 1999; Ramos, Sedivi, & Sweet, 1998; Sproull, 1986). These surveys are fast and require little technological skill to develop as they are displayed in a basic-text format. Respondents are asked to reply to the email and indicate their responses in the reply message or as part of the attached file. These surveys require little technological skill on the part of the respondent, but researchers (Couper, Blair, & Triplett, 1997; Tse, et., al., 1995; Schaeffer & Dillman, 1998) have found that respondents experience some difficulties such as remembering they must reply to the message before answering the survey questions and having trouble converting an attachment. Additionally, these surveys raise concerns regarding privacy and anonymity as the respondent's e-mail address is generally included with his/her responses.

A third type of electronic survey is posted on the World Wide Web (WWW). Respondents are usually sent an e-mail message with a link to the URL address for the survey. Web-based surveys can be designed to include a wide variety of response options (e.g., check boxes, Likert scales, pull-down menus) as well as skip patterns, graphics and sound (Bowers, 1999; Bradley, 1999; Dillman, 2000; Watt, 1997). These surveys also offer great advantages in terms of data analysis as responses can easily be downloaded into a spreadsheet or statistical analysis software program, but respondents should also be concerned with the privacy as their responses are transferred over the WWW. Of the three types of electronic surveys we just discussed, these surveys require the greatest amount of technological knowledge and skill of the researcher(s) and respondents.

Due to the technological knowledge and skill required to develop electronic surveys, especially web-based surveys, the leadership in terms of their development has come in large part from technology specialists or individuals with a background in technology. Survey methodology professionals have not been the driving force behind the use of electronic surveys. The challenge for survey methodologists is to tailor sound principles of survey design and implementation to the use of electronic surveys (Dillman, 2000; Dillman & Bowker, 2000). However, to harness the potential of using the Internet for designing and implementing surveys, professionals knowledgeable about survey methodology must provide a more visible presence. Are survey professionals ready to accept and use electronic surveys as part of their methodological repertoire? Before electronic surveys are widely accepted and used on a regular basis, input must be gathered from survey professionals.

Purpose

The purpose of this study was to gather the perceptions and recommendations of survey researchers regarding the use of electronic surveys. These researchers were asked to respond to specific issues that pertain to the use of electronic surveys. In addition, these researchers were asked to describe conditions under which the use of e-mail or web-based surveys would be most appropriate, define appropriate samples, identify the major weaknesses, and offer recommendations for other researchers that plan to use email or the Internet to assist their survey research projects.

Methods

Instrumentation

The survey instrument consisted of three sections. First, a four-point Likert-scale instrument was developed to address issues regarding the use of electronic mail or the Internet in survey research. These items were written to reflect issues such as sampling frame, privacy, technology, and response rate raised in the literature discussed earlier. The second section consisted of four open-ended questions to solicit feedback regarding the uses of electronic surveys in survey research, the limitations of such surveys, the types of samples for which such surveys would be appropriate, and suggestions for those interested in using electronic mail or the Internet for survey research. Finally, the third section was included to gather information about the participants in this study. Items in this section specifically addressed participant's background and confidence in using technology (i.e., electronic mail and the Internet), their current professional position, and their involvement in their profession.

Procedures

The participants were identified on a published membership list of the Survey Research SIG from the American Educational Research Association (AERA). This list was obtained from and used with the permission of the Director of the Survey Research Special Interest Group. This list included 163 members for which complete mailing information was available. Each subject received a packet that included the survey instrument and a postcard. In order to assure anonymity, they were asked to return the postcard separately indicating whether they responded to the survey. A total of 63 responses were received. An additional 35 surveys were returned as undeliverable as members may have changed their place of employment or retired. After subtracting these 35 from the overall sample, a response rate of 49% was obtained (i.e., 63 out of 128). A total of 64 postcards were received. Of these 64, 56 indicated that they returned the survey and 8 stated that they did not return the survey. Three reasons were expressed from the group of eight non-responding individuals. Three (3) indicated that they were just too busy, 3 indicated that they were no longer active in survey research, and 2 indicated that they were retired.

Sample

The majority of these respondents (53%) were employed at a college or university. An additional 13% were working as consultants while 10% worked for testing organizations, 8% for school systems, and 8% for research and development organizations. The remaining 8% were employed by state or federal agencies or private industry. Respondents indicated a wide range of years in their current position, from 1 to 30 years, with an average of 13.23 years. The number of years in their profession ranged from 1 to 45, with an average of 17.7 years. Membership in AERA ranged from 1 year to 35, with an average of 12.1 years. Forty-three percent of the respondents identified AERA as their primary professional organization and had been AERA members for an average of 15.2 years.

Results

Use of Electronic Mail and the Internet

Overall, the sample participants reported frequent use and a high level of confidence in using electronic mail and the Internet. Ninety (90) percent reported using email everyday and 57% described themselves as everyday Internet users, with 78% reporting use of the Internet at least 5 days per week. Participants were also asked to describe their confidence in using electronic mail and the Internet. In general, they reported being very confident in their ability to use email (e.g., composing and responding to messages, sending messages to more than one person and sending attachments). They were also confident in their ability to use the Internet to do things like find a web address, use a search engine, and download information. The only area in which these participants expressed a concern was creating and maintaining a web page.

General Perceptions of Electronic Surveys

Each participant was asked to respond to 33 Likert-scale items pertaining to the use of email or web-based surveys. Six of these items were reverse-coded so that a higher score would consistently reflect a more favorable attitude toward the use of email or web-based surveys. Internal consistency reliability (Cronbach's alpha) was estimated at .83. Overall, participants responded favorably to statements regarding the use of email or web-based surveys. Table 1 provides a summary of means, standard deviations, and frequencies for the survey items. These items are displayed in descending order by mean response.

Table 1
Summary of Perceptions of Electronic Surveys

Survey Item	**N^a**	**Mean (SD)**	**Strongly Disagree Or Disagree N (%)**	**Strongly Agree Or Agree N (%)**
Electronic surveys reduce research costs. (e.g., postage, phone)	60	3.42 (.56)	2 (3.3%)	58 (96.7%)
Respondents to electronic surveys would be more comfortable with technology than non-respondents	62	3.32 (.59)	4 (6.5%)	58 (93.5%)
Electronic mail messages would be an effective way to pre-notify individuals regarding a survey they are about to receive	61	3.28 (.61)	3 (4.9%)	58 (95.1%)
Researchers would use electronic surveys if they yielded data ready to be imported into a statistical analysis program such as SAS or SPSS.	59	3.12 (.70)	9 (15.3%)	50 (84.7%)
Electronic mail messages would be effective as a follow-up technique to encourage response to a mail survey.	61	3.12 (.61)	6 (9.8%)	55 (90.2%)
I have considered the use of electronic mail or Internet in my research.	61	3.03 (.60)	8 (13.1%)	53 (86.9%)
I would respond to a web-based survey if I simply had to click on the URL address the researcher placed in an e-mail message.	61	3.02 (.62)	9 (14.7%)	52 (85.3%)
Electronic surveys will be returned more rapidly than traditional pencil-and-paper surveys.	61	2.98 (.76)	12 (19.7%)	49 (80.3%)
Individuals would respond to a web-based survey if they simply had to click on the URL address the researcher placed in an e-mail message.	59	2.98 (.51)	8 (13.6%)	51 (86.4%)
Electronic surveys reduce the time and labor required to prepare data for analysis.	59	2.97 (.69)	13 (22.0%)	46 (78.0%)

Electronic surveys eliminate the need to transcribe responses to open-ended questions.	60	2.95 (.77)	15 (25.0%)	45 (75.0%)
Electronic surveys should allow for text editing capabilities	57	2.95 (.72)	12 (21.1%)	45 (78.9%)
Electronic surveys would be useful for alumni surveys.	57	2.95 (.66)	12 (21.1%)	45 (78.9%)
E-mail surveys would require too much time and effort for respondents.	61	2.90 (.37)	53 (86.9%)	8 (13.1%)
I would access a web page to respond to a survey that interested me.	61	2.89 (.71)	15 (24.6%)	46 (75.4%)
In general, people would access a web page to respond to a survey if the topic was of interest.	58	2.85 (.56)	14 (24.1%)	44 (75.9%)
I would use electronic surveys if responses could be directly imported into a file for data analysis.	56	2.79 (.62)	16 (28.6%)	40 (71.4%)
Electronic surveys and pencil-and-paper surveys yield comparable information.	52	2.72 (.57)	14 (26.9%)	38 (73.1%)
The use of electronic surveys would make it more difficult to obtain Institutional Review Board (IRB) approval.	48	2.60 (.75)	33 (68.8%)	15 (31.2%)
Potential respondents would find electronic surveys more interesting than pencil-and-paper surveys.	60	2.50 (.57)	32 (53.3%)	28 (46.7%)
People would not respond to electronic surveys because they would get lost along with junk mail received from listservs and newsgroups.	56	2.50 (.57)	28 (50.0%)	28 (50.0%)
Electronic surveys are better suited for an Internet web page compared to being included as part of an e-mail message.	58	2.48 (.57)	32 (55.2%)	26 (44.8%)
Electronic surveys would be useful for political polls.	59	2.48 (.94)	27 (45.8%)	32 (54.2%)
In general, people prefer hard copies of surveys.	53	2.45 (.67)	27 (50.9%)	26 (49.1%)
The reliability of electronic surveys is equal to or stronger than that estimated for paper-and-pencil surveys.	51	2.45 (.67)	25 (49.0%)	26 (51.0%)
In general, I would expect a greater response to electronic surveys.	60	2.43 (.75)	31 (51.7%)	29 (48.3%)
Using an electronic survey would communicate more urgency than traditional mail surveys	61	2.41 (.59)	37 (60.7%)	24 (39.3%)
I would be more likely to respond to a an electronic survey than a pencil-and-paper survey.	60	2.40 (.81)	38 (63.3%)	22 (36.7%)
Individuals would not respond to electronic surveys because of issues related to anonymity.	57	2.39 (.68)	24 (42.1%)	33 (57.9%)
In general, individuals would be more likely to respond to an electronic survey.	56	2.36 (.62)	34 (60.7%)	22 (39.3%)
Electronic surveys do not allow for anonymity.	60	2.30 (.83)	26 (43.3%)	34 (56.7%)
Respondents would complete more items on an electronic survey compared to a pencil-and-paper survey.	61	2.23 (.62)	43 (70.5%)	18 (29.5%)
Responses to electronic surveys would be less likely to be influenced by social desirability compared to traditional paper surveys.	59	2.22 (.59)	45 (76.3%)	14 (23.7%)
People would make fewer mistakes when responding to questions in electronic surveys.	59	2.17 (.46)	47 (79.7%)	12 (20.3%)
Receiving a survey through e-mail would be more personalized than through traditional mail.	62	2.11 (.55)	49 (79.0%)	13 (21.0%)
NOTE: Response scale – 1=Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree				

These survey professionals were most positive in terms of the reduction of costs (i.e., postage, phone charges) associated with electronic surveys, the use of electronic mail for pre-notification or follow-up purposes as a complement to other survey delivery methods, and the compatibility of data with existing software programs. They also indicated that the lack of a tangible reward would not prevent individuals from responding and that they would respond to a web-based survey if all they had to do was click on the HTML address from an email message.

The bulk of the less favorable responses pertained to respondents' knowledge and experience with technology. They believed that individuals who were not comfortable with technology would not respond. In addition, they indicated that electronic surveys are less personalized than traditional mail surveys, people will make more mistakes when responding, their responses will be influenced by issues of social desirability, and they will not complete as many items as they might have in a pencil-and-paper survey. Finally, these survey researchers expressed a need for passwords to access web-based surveys, a concern that respondents would not be as likely to respond to sensitive issues, or not respond at all due to a concern for their anonymity.

There were also a few areas in which these survey professional were very uncertain. In other words, they were very balanced in terms of their agreement and disagreement regarding several items. These items regarded the comparability of response rate and reliability estimates for electronic and mail surveys, the extent to which people prefer hard copies of surveys or find electronic surveys more interesting, and the appropriateness of listserves as a sampling source for electronic surveys.

Advice from survey professionals

In addition to general perceptions, specific advice was solicited regarding the most effective use of electronic surveys, appropriate samples, limitations, and recommendations for researchers considering the use of electronic surveys. This advice was gathered using four open-ended questions.

Effective use of electronic surveys. Thirty-seven (37) respondents provided guidance regarding the effective use of electronic surveys in survey research. Nearly half (48%, n=18) of the respondents indicated that such surveys could be used most effectively for targeted populations such as professional or business groups with published email addresses or as "in-house" surveys. Twenty-seven percent (n=10) simply indicated that email or web-based surveys would be more efficient, obtaining responses faster and produce data that could be directly stored in a format suitable for analysis and 16% (n=6) described specific uses of email or web-based surveys, including pre-notification of subjects, follow-up of mail surveys, marketing research, needs assessments, and longitudinal studies. The remaining three respondents indicated that such surveys must be carried out under specific conditions, keeping the surveys short and simple to respond to and have some mechanism such as a password to maintain anonymity.

Appropriate Samples for Electronic Surveys. A total of 35 respondents offered recommendations regarding samples that would be appropriate for electronic surveys. These suggestions primarily focused on samples that have access to and the ability to use technology. The majority of these professionals' responses (n=32, 91.5%) described specific types of groups that have access to technology. Specific samples identified included listservs, professional memberships, alumni groups, "in house" employee groups, and University professors. The remaining three respondents simply indicated that samples had to be small and clearly defined.

Limitations of electronic surveys. Forty-eight (48) participants offered comments regarding the limitation of email or web-based surveys. The majority (n=25, 52%) of these responses described sampling limitations. More specifically, these sampling concerns pertained to the restricted nature of such samples in that respondents must have access to and be comfortable using technology and that such samples would not accurately represent the general population.

A second concern expressed regarded confidentiality and a lack of privacy, expressed by 15 respondents (31.3%).

Concerns were voiced that the invitation to respond to email or web-based surveys might be perceived as junk mail and mass mailings to published email lists might be perceived as "spam." Furthermore, there were concerns regarding the security of the information posted and submitted through email or web-based surveys, raising questions about the invasion of the privacy of respondents and security of information on the Internet. Several researchers used the phase "Big Brother" to describe their concern with privacy of information.

A third group of concerns (n=12, 25%) pertained to the credibility and authenticity of the results from electronic surveys. Many of these surveys are open to responses from individuals outside the targeted sample. Specific recommendations were made to have safeguards in place to verify the authenticity of respondents. Such safeguards might take the form of passwords that only allow those who were invited to complete the survey. Without such safeguards, the credibility of the data received from respondents is questionable.

A final group of limitations (n=6, 12.5%) were methodological in nature. Such surveys require a great deal of time and technological skill to develop and implement. Several respondents raised questions about the compatibility with traditional pencil and paper surveys, commenting on the difficulty in formatting surveys to fit in web pages and the limited number of incentives that could be provided for potential respondents.

Suggestions for Others Interested in Using Electronic Surveys. Finally, 23 respondents made suggestions for others. These suggestions primarily regarded issues of sampling, survey format, and procedures. Ten suggestions (43.5%) made reference to sampling issues. Specifically, five recommendations were made to pre-sample the population to determine their interest in participating. The remaining sample-related comments were offered as cautions to the survey researcher in that he/she should be aware that the sample will be limited and that technology will not be uniform among members of the sample.

Eight respondents (34.8%) made recommendations regarding design and format. Three recommended a simple, short survey and three simply advised that close attention be paid to sound survey design principles while the remaining two specifically indicated a preference for graphically-pleasing web-based surveys.

The remaining five suggestions (21.7%) were categorized as procedural. Two respondents recommended that the time is now to use electronic surveys, before such surveys become too common. Another researcher suggested that respondents be given an option to respond using a hard copy while one recommended the use of email as a follow-up technique. The final comment simply stated 'be skeptical."

Discussion and Recommendations

Consistent with prior literature (Bowers, 1999; Crawford, Couper & Lamias, 2001; Eaton, 1997; Kaye & Johnson, 1999; Kiessler & Sproull, 1986; Weissbach, 1997), we found that the primary concerns expressed by survey professionals in this study regarded sampling issues. These concerns regarded sample's access and ability to use the required technology, their authenticity and their privacy. Advice from this group of professionals specifically focused on the recognition of limitations of electronic survey samples and precautions that should be taken to establish credible samples and protect respondents' privacy.

First of all, it is clear that the sampling frame is still somewhat limited when using electronic surveys and survey professionals must acknowledge these limitations when conducting their research. Samples with access to the Internet have not typically represented the general population (GVU, 1998; Sheehan & Hoy, 1999). For this reason, professional or business groups with published e-mail addresses have often been targeted as samples. However, the Internet is exploding and becoming increasing more accessible to the general population as approximately 41.5% of US households now have access, an increase of 58% in less than two years (Department of Commerce, 2000). Access is still more frequent among those who live in urban areas, with higher incomes and higher levels of education. However, the most rapid increases in access are occurring in rural areas, among individuals with some college experience, and individuals over 50 (Department of Commerce, 2000). Such increases will continue and the gaps between Internet users and the general population will continue to close. The increase in Internet access and reliable e-mail addresses will allow for a greater range in samples for future electronic surveys.

Researchers must also recognize that samples will vary a great deal in terms of their technological capability, both in terms of equipment and respondent knowledge and skill. This variation must be kept in mind when designing electronic surveys. Although web-based surveys allow for much more innovative features than plain text e-mail surveys, respondents may have difficulty accessing the survey and will not be able to respond. Furthermore, most people are not accustomed to the process used to respond to an electronic survey (e.g., selecting from a pull-down menu, clicking a radial button, scrolling from screen to screen) and will need specific instructions that guide them through each questions and the manner in which they should respond.

Based on the advice of survey professionals, we recommend that samples be pre-notified using an e-mail message to determine the technological capacity of the sample and their willingness to participate in the study. This will help ensure that the survey will be accessible to members in the sample and help prevent the perceptions of "spamming" that might occur due to continued unsolicited e-mail messages (Mehta & Sivadas, 1995; Sheehan & Hoy, 1999). This communication should be personalized and provide for the essential elements of mailed cover letters, including provide a clear overview of the study's purpose, motivation to respond, assurances of confidentiality and privacy and who they contact should they have questions. This advice was reinforced by a recent meta-analysis of electronic survey studies which found personalized pre-notification and number of contacts to influence response rate (Cook, Heath, & Thompson, 2000).

Once samples are identified and pre-notified, they need to be protected in terms of their authenticity, confidentiality, and privacy. Measures should be taken to reduce sampling error. Access to web-based surveys must be limited to the targeted sample. Unrestricted sample surveys that allow anyone access are unacceptable. Whereas many unscientific online polls boast large samples, there is often little or no attempt to ensure the quality and validity of such samples.

Samples must be clearly defined and authenticated. Researchers should consider using passwords or PIN numbers to control for sampling error and establish credible samples (Bowers, 1999; Bradley, 1999; Dillman, Tortora, & Bowker, 1998). In the case that passwords or PIN numbers are not used, responding samples should be carefully examined and those that are not eligible should be eliminated to maintain consistency with the sampling plan and yield credible results.

Additional precautions must be taken to protect respondents' privacy and ensure the confidentiality of their responses. Several researchers have experienced negative feedback from respondents regarding privacy issues (Couper, Blair, & Triplett, 1997; Mehta & Sivadas, 1995; Sheehan & Hoy, 1999). In analyzing server logs from electronic surveys, Jeavons (1998) found that individuals stopped completing surveys when their email address was requested. Respondents must feel comfortable when responding to electronic surveys and trust researchers have taken precautions to guard their privacy. Minimally, researchers should make assurances of confidentiality in the pre-notification e-mail (Couper, Blair, & Triplett, 1997; Kieslerr & Sproull, 1986; Schaeffer & Dillman, 1998). Further protection of respondents' privacy can be provided by separating e-mail addresses upon receipt of the completed surveys (Sheehan & Hoy, 1999) or programming the return to include the researcher's e-mail address, not that of the respondent (Shannon & Bradshaw, 2000). Using secure servers and encryption methods should also be employed as an additional protection of respondents' privacy

In conclusion, electronic surveys web-based must utilize principles of sound survey design. Research studies must also focus on the adaptability of such principles for electronic survey formats so that survey professionals can take full advantage of the benefits of such surveys without sacrificing the integrity of their data and placing respondents at risk in terms of confidentiality and privacy. As methods pertaining to the design and implementation of electronic surveys are refined, they will be used more frequently to conduct scholarly research. This also means that Institutional Review Boards (IRB) will encounter increasing numbers of proposals and the issues of confidentiality and privacy will become increasingly important and policies pertaining to the protection of human subjects as participants in electronic surveys and other types of research using the Internet will need to be developed.

References

Babbie, E. (1990). *Survey Research Methods* (2nd ed.). Belmont, CA: Wadsworth.

Baruch, Y. (1999). Response rates in academic studies: A comparative analysis. *Human Relations, 52,* 421-434.

Bowers, D. K. (1999). FAQs on online research. *Marketing Research, 10*(1), 45-48.

Bradley, N. (1999). Sampling for Internet surveys: An examination of respondent selection for Internet research. *Journal of the Market Research Society, 41*(4), 387-395.

Cook, C., Heath, F., & Thomson, R. (2000). A meta-analysis of response rates in web- or Internet-based surveys. *Educational & Psychological Measurement*, 60(6), 821-826.

Couper, M. P. Blair, J. & Triplett, T. (1997). *A comparison of mail versus email for surveys of employees in federal statistical agencies.* Paper presented at the annual meeting of the American Association for Public Opinion Research, Norfolk, VA.

Couper, M. P. & Nichols, W. L. (1998). The history and development of computer assisted survey information collection methods. In M. P. Couper, R. P. Baker, J. Bethlehem, C.Z.E. Clark, J. Martin, W.L. Nichols, & J. M. O'Reilly (Eds.). *Computer assisted survey information collection* (pp. 1-22). New York: John Wiley & Sons, Inc.

Crawford, S. D., Couper, M. P. & Lamias, M. J. (2001). Web surveys: Perception of burden. *Social Science Computer Review, 19*, 146-162.

Department of Commerce (2000, October). *Falling through the net: Toward digital inclusion.* Washington, DC: Author.

Dillman, D. A. (2000). *Mail and Internet surveys: The tailored design method.* New York: John Wiley and Sons, Inc.

Dillman, D. A. (1978).). *Mail and telephone surveys: The total design method.* New York: John Wiley and Sons, Inc.

Dillman, D. A. & Bowker, D. K. (2000). *The web questionnaire challenge to survey methodologists.* [Online]. Available: http://sesrc.wsu.edu/dillman/papers.htm.

Dillman, D. A., Tortora, R. D., Bowker, D. (1998). *Principles for constructing web surveys: An initial statement.* (Technical Report No. 98-50). Pullman, WA: Washington State University Social and Economic Sciences Research Center.

Eaton, B. (1997). Internet surveys: Does WWW stand for "why waste the work?" *Marketing Research*

Review, June/July, Article 0244. Available:http://www.Quirks.com

Fowler, F. J. (1993). *Survey Research Methods* (2nd ed.) Newbury Park: Sage Publications.

GVU's 10th WWW User Survey (October, 1998). General Demographic Summary [On-line]. Available http://www.gvu.gatech.edu/user_surveys/survey-1998-10/reports/

Herberlein, T. A. & Baumgartner, R. (1978). Factors affecting response rates to mailed questionnaires: A quantitative analysis of the published literature. *American Sociological Review, 43*, 447-462.

Jeavons, A. Ethology and the Web: Observing respondent behavior in Web surveys. Proceedings of the Worldwide Internet Conference, Amsterdam: ESOMAR, 1998. Available: http://w3.one.net/~andrewje/ethology.html.

Kaye, B. K. & Johnson, T. J. (1999). Research methodology: Taming the cyber frontier. *Social Science Computer Review*, 17, 323-337.

Kiesler, S. & Sproull, L. S. (1986). Responses effects in the electronic survey. *Public Opinion Quarterly, 50*, 402-413.

Lavrakas, P. J. (1993). *Telephone Survey Methods: Sampling, Selection, and Supervision* (2nd ed.) Newbury Park: Sage Publications.

Linsky, A. S. (1975). Stimulating responses to mailed questionnaires: A review. *Public Opinion Quarterly, 39*, 82-101.

Mehta, R. & Sivadas, E. (1995). Comparing response rates and responses content in mail versus electronic mail surveys. *Journal of the Market Research Society, 37*(4), 429-439.

Ramos, M., Sedivi, B. M., & Sweet, E. M. (1998). Computerixed self-adnministered questionnaires. . In M. P. Couper, R. P. Baker, J. Bethlehem, C.Z.E. Clark, J. Martin, W.L. Nichols, & J. M. O'Reilly (Eds.). *Computer assisted survey information collection* (pp. 389-408). New York: John Wiley & Sons, Inc.

Schaeffer, D. R. & Dillman, D. A. (1998). Development of standard e-mail methodology: Results on an experiment. Public Opinion Quarterly,62(3), 378-397.

Shannon, D. M. & Bradshaw, C. C. (2002). A comparison of response rate, speed and costs of mail and electronic surveys. *Journal of Experimental Education,* 70(2), in press.

Sheehan, K. B. & Hoy, M. G. (1999). Using e-mail to survey Internet users in the United States: Methodology and Assessment. *Journal of Computer Mediated Communication, 4*(3). Available: http://www.ascusc.org/jcmc/vol4/issue3/sheehan.html.

Solomon, David J. (2001). Conducting web-based surveys. *Practical Assessment, Research & Evaluation, 7* (19). Available online: http://ericae.net/pare/getvn.asp?v=7&n=19.

Sproull, L. S. (1986). Using electronic mail for data collection in organizational research. *Academy of Management Journal*, *29*(1), 156-169.

Sudman, S. & Bradburn, N. M. (1982). *Asking Questions: A practical guide to questionnaire design.* San Francisco: Jossey-Bass Publishers.

Tse, A. C. B., Tse, K. C., Yin, C. H., Ting, C. B., Yi, K. W., Yee, K. P., & Hong, W. C. (1995). Comparing two methods of sending out questionnaires: E-mail versus snail mail. *Journal of the Market Research Society, 37*(4), 441-446.

Watt, J. H. (1997). Using the Internet for quantitative survey research. *Marketing Research Review,* June, Article 0248. Available: http://www.Quirks.com

Weissbach, S. (1997) Internet Research: Still a few hurdles to clear. *Marketing Research Review,* June/July. Article 0249. Available: http://www.Quirks.com

Yu, J. & Cooper, H. (1983). A quantitative review of research deign effects on response rates to questionnaires. *Journal of Marketing Research, 20*(1), 36-44.

Contact

Please direct all correspondence to the first author at:

David Shannon
4036 Haley Center – EFLT
Auburn University
Auburn, Alabama 36839-5221

(334) 844-3071, FAX: (334) 844-3072
shanndm@auburn.edu

Descriptors: *World Wide Web; *Survey Methods; Response Rates [Questionnaires]; *Surveys; Electronic Mail

ADODB.Recordset error '800a0e78'

Operation is not allowed when the object is closed.

/pare/getvn.asp, line 179

Home | Articles | Subscribe | Review | Policies

Volume: 8 7 6 5 4 3 2 1

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ☐ Go

Copyright 2002, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Osborne, Jason & Elaine Waters (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research & Evaluation*, 8(2). Available online: http://ericae.net/pare/getvn.asp?v=8&n=2. This paper has been viewed 10598 times since 1/7/02.

▸ Find similar papers in
ERICAE Full Text Lil
Pract Assess, Res &
ERIC RIE & CIJE 19
ERIC On-Demand D

▸ Find articles in ERIC w
Osborne, Jason
Elaine Waters

Four Assumptions Of Multiple Regression That Researchers Should Always Test

Jason W. Osborne and Elaine Waters
North Carolina State University and University of Oklahoma

Most statistical tests rely upon certain assumptions about the variables used in the analysis. When these assumptions are r the results may not be trustworthy, resulting in a Type I or Type II error, or over- or under-estimation of significance or ef (s). As Pedhazur (1997, p. 33) notes, "Knowledge and understanding of the situations when violations of assumptions lea serious biases, and when they are of little consequence, are essential to meaningful data analysis". However, as Osborne, Christensen, and Gunter (2001) observe, few articles report having tested assumptions of the statistical tests they rely on f drawing their conclusions. This creates a situation where we have a rich literature in education and social science, but we forced to call into question the validity of many of these results, conclusions, and assertions, as we have no idea whether tl assumptions of the statistical tests were met. Our goal for this paper is to present a discussion of the assumptions of multi regression tailored toward the practicing researcher.

Several assumptions of multiple regression are "robust" to violation (e.g., normal distribution of errors), and others are ful the proper design of a study (e.g., independence of observations). Therefore, we will focus on the assumptions of multipl regression that are not robust to violation, and that researchers can deal with if violated. Specifically, we will discuss the assumptions of linearity, reliability of measurement, homoscedasticity, and normality.

VARIABLES ARE NORMALLY DISTRIBUTED.

Regression assumes that variables have normal distributions. Non-normally distributed variables (highly skewed or kurto variables, or variables with substantial outliers) can distort relationships and significance tests. There are several pieces o information that are useful to the researcher in testing this assumption: visual inspection of data plots, skew, kurtosis, and plots give researchers information about normality, and Kolmogorov-Smirnov tests provide inferential statistics on norma

Outliers can be identified either through visual inspection of histograms or frequency distributions, or by converting data t scores.

Bivariate/multivariate data cleaning can also be important (Tabachnick & Fidell, p 139) in multiple regression. Most regr multivariate statistics texts (e.g., Pedhazur, 1997; Tabachnick & Fidell, 2000) discuss the examination of standardized or studentized residuals, or indices of leverage. Analyses by Osborne (2001) show that removal of univariate and bivariate o can reduce the probability of Type I and Type II errors, and improve accuracy of estimates.

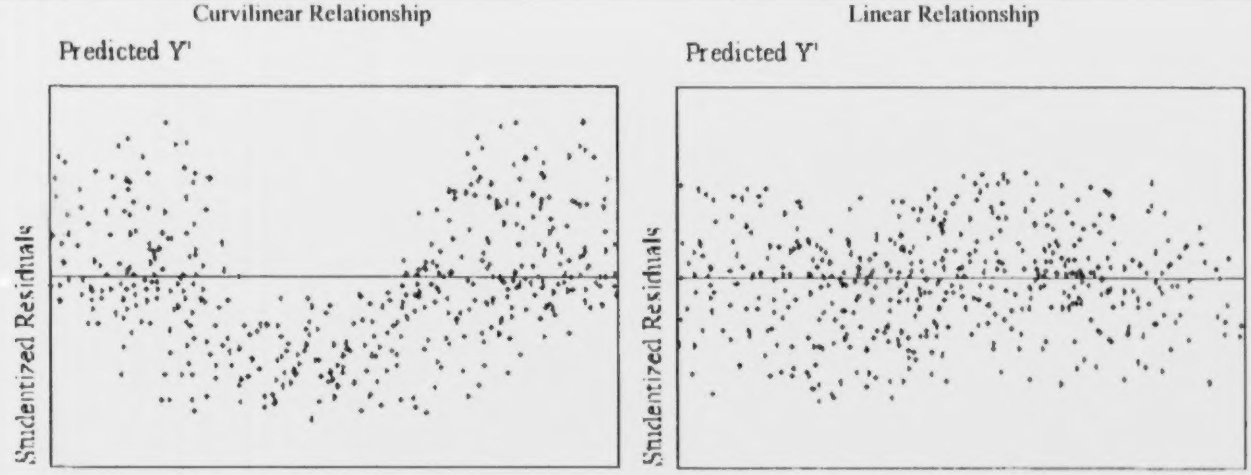
Outlier (univariate or bivariate) removal is straightforward in most statistical software. However, it is not always desirabl remove outliers. In this case transformations (e.g., square root, log, or inverse), can improve normality, but complicate th interpretation of the results, and should be used deliberately and in an informed manner. A full treatment of transformatio beyond the scope of this article, but is discussed in many popular statistical textbooks.

ASSUMPTION OF A LINEAR RELATIONSHIP BETWEEN THE INDEPENDENT AND DEPENDENT VARIAE

Standard multiple regression can only accurately estimate the relationship between dependent and independent variables i relationships are linear in nature. As there are many instances in the social sciences where non-linear relationships occur (anxiety), it is essential to examine analyses for non-linearity. If the relationship between independent variables (IV) and t dependent variable (DV) is not linear, the results of the regression analysis will *under-estimate* the true relationship. This estimation carries two risks: increased chance of a Type II error for that IV, and in the case of multiple regression, an incr risk of Type I errors (over-estimation) for other IVs that share variance with that IV.

Authors such as Pedhazur (1997), Cohen and Cohen (1983), and Berry and Feldman (1985) suggest three primary ways tc non-linearity. The first method is the use of theory or previous research to inform current analyses. However, as many pr researchers have probably overlooked the possibility of non-linear relationships, this method is not foolproof. A preferabl of detection is examination of residual plots (plots of the standardized residuals as a function of standardized predicted val readily available in most statistical software). Figure 1 shows scatterplots of residuals that indicate curvilinear and linear relationships.

Figure 1. Example of curvilinear and linear relationships with standardized residuals by standardized predicted values.



The third method of detecting curvilinearity is to routinely run regression analyses that incorporate curvilinear component (squared and cubic terms; see Goldfeld and Quandt, 1976 or most regression texts for details on how to do this) or utilizin nonlinear regression option available in many statistical packages. It is important that the nonlinear aspects of the relation accounted for in order to best assess the relationship between variables.

VARIABLES ARE MEASURED WITHOUT ERROR (RELIABLY)

The nature of our educational and social science research means that many variables we are interested in are also difficult measure, making measurement error a particular concern. In simple correlation and regression, unreliable measurement c relationships to be *under-estimated* increasing the risk of Type II errors. In the case of multiple regression or partial corre effect sizes of other variables can be *over-estimated* if the covariate is not reliably measured, as the full effect of the covar would not be removed. This is a significant concern if the goal of research is to accurately model the "real" relationships in the population. Although most authors assume that reliability estimates (Cronbach alphas) of .7-.8 are acceptable (e.g., Nunnally, 1978) and Osborne, Christensen, and Gunter (2001) reported that the average alpha reported in top Educational Psychology journals was .83, measurement of this quality still contains enough measurement error to make correction wo as illustrated below.

Correction for low reliability is simple, and widely disseminated in most texts on regression, but rarely seen in the literatu argue that authors should correct for low reliability to obtain a more accurate picture of the "true" relationship in the popu and, in the case of multiple regression or partial correlation, to avoid over-estimating the effect of another variable.

Reliability and simple regression

Since "the presence of measurement errors in behavioral research is the rule rather than the exception" and "reliabilities o measures used in the behavioral sciences are, at best, moderate" (Pedhazur, 1997, p. 172); it is important that researchers of accepted methods of dealing with this issue. For simple regression, Equation #1 provides an estimate of the "true" rela between the IV and DV in the population:

$$\overset{*}{r}_{12} = \frac{r_{12}}{\sqrt{r_{11}r_{22}}} \tag{1}$$

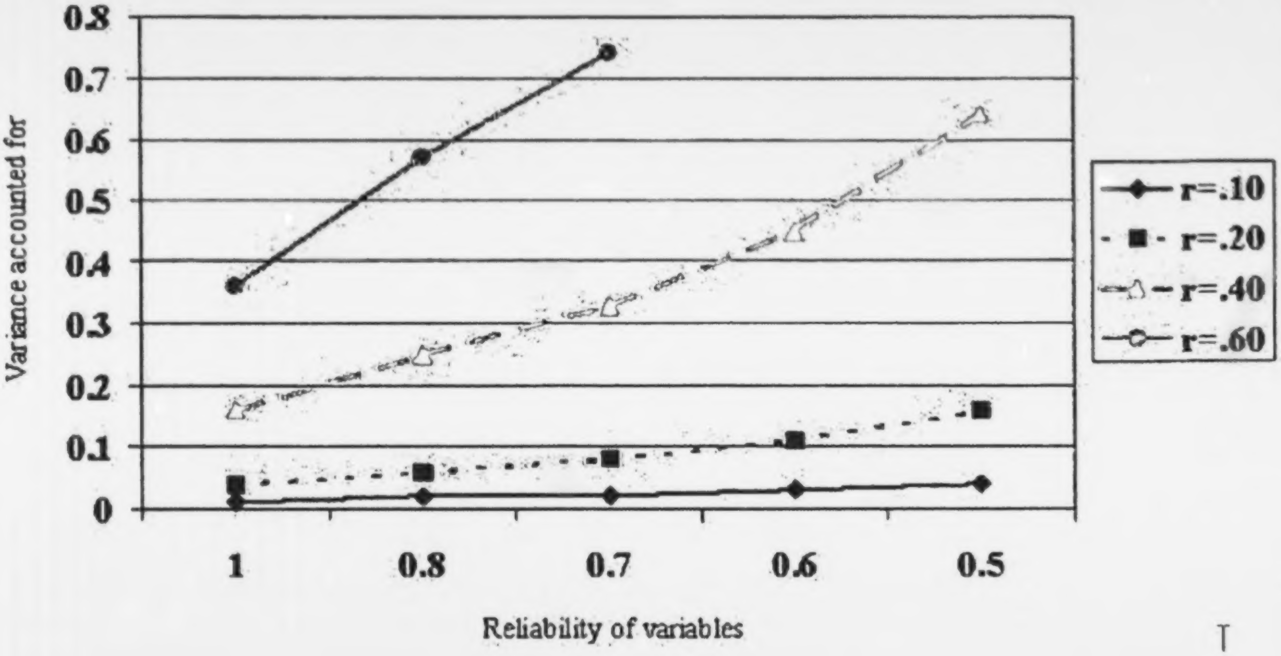
In this equation, r_{12} is the observed correlation, and r_{11} and r_{22} are the reliability estimates of the variables. Table 1 and presents examples of the results of such a correction.

Table 1: *Values of r and r^2 after correction for attenuation*

	Reliability of DV and IV									
	Perfect measurement		.80		.70		.60		.50	
Observed *r*	*r*	r^2	*r*	r^2	*r*	r^2	*r*	r^2	*r*	r^2
.10	.10	.01	.13	.02	.14	.02	.17	.03	.20	.04
.20	.20	.04	.25	.06	.29	.08	.33	.11	.40	.16
.40	.40	.16	.50	.25	.57	.33	.67	.45	.80	.64
.60	.60	.36	.75	.57	.86	.74	--	--	--	--

Note: for simplicity we show an example where both IV and DV have identical reliability estimates. In some of these hypothetical examples we would produce impossible values, and so do not report these.

Figure 2: Change in variance accounted for as correlations are corrected for low reliability



As Table 1 illustrates, even in cases where reliability is .80, correction for attenuation substantially changes the effect siz (increasing variance accounted for by about 50%). When reliability drops to .70 or below this correction yields a substant different picture of the "true" nature of the relationship, and potentially avoids a Type II error.

Reliability and Multiple Regression

With each independent variable added to the regression equation, the effects of less than perfect reliability on the strength relationship becomes more complex and the results of the analysis more questionable. With the addition of one independ variable with less than perfect reliability each succeeding variable entered has the opportunity to claim part of the error va left over by the unreliable variable(s). The apportionment of the explained variance among the independent variables will incorrect. The more independent variables added to the equation with low levels of reliability the greater the likelihood th variance accounted for is not apportioned correctly. This can lead to erroneous findings and increased potential for Type for the variables with poor reliability, and Type I errors for the other variables in the equation. Obviously, this gets increa complex as the number of variables in the equation grows.

A simple example, drawing heavily from Pedhazur (1997), is a case where one is attempting to assess the relationship bet variables controlling for a third variable ($r_{12.3}$). When one is correcting for low reliability in all three variables Equation used:

$$r^{*}_{123} = \frac{r_{33}r_{12} - r_{13}r_{23}}{\sqrt{r_{11}r_{33} - r_{13}^{2}}\sqrt{r_{22}r_{33} - r_{23}^{2}}} \qquad (2)$$

Where r_{11}, r_{22}, and r_{33} are reliabilities, and r_{12}, r_{23}, and r_{13} are relationships between variables. If one is only correcting reliability in the covariate one could use Equation #3:

$$r^{*}_{12.3} = \frac{r_{33}r_{12} - r_{13}r_{23}}{\sqrt{r_{33} - r_{13}^{2}}\sqrt{r_{33} - r_{23}^{2}}} \tag{3}$$

Table 2 presents some examples of corrections for low reliability in the covariate (only) and in all three variables.

Table 2: *Values of* $r_{12.3}$ *and* $r^2_{12.3}$ *after correction low reliability*

Examples:				Reliability of Covariate			Reliability of All Variables		
				.80	.70	.60	.80	.70	.60
r_{12}	r_{13}	r_{23}	Observed $r_{12.3}$	$r_{12.3}$	$r_{12.3}$	$r_{12.3}$	$r_{12.3}$	$r_{12.3}$	$r_{12.3}$
.3	.3	.3	.23	.21	.20	.18	.27	.30	.33
.5	.5	.5	.33	.27	.22	.14	.38	.42	.45
.7	.7	.7	.41	.23	.00	-.64	.47	.00	--
.7	.3	.3	.67	.66	.65	.64	.85	.99	--
.3	.5	.5	.07	-.02	-.09	-.20	-.03	-.17	-.64
.5	.1	.7	.61	.66	.74	.90	--	--	--

Note: In some examples we would produce impossible values that we do not report.

Table 2 shows some of the many possible combinations of reliabilities, correlations, and the effects of correcting for only covariate or all variables. Some points of interest: (a) as in Table 1, even small correlations see substantial effect size (r^2) when corrected for low reliability, in this case often toward reduced effect sizes (b) in some cases the corrected correlatior only substantially different in magnitude, but also in direction of the relationship, and (c) as expected, the most dramatic c occur when the covariate has a substantial relationship with the other variables.

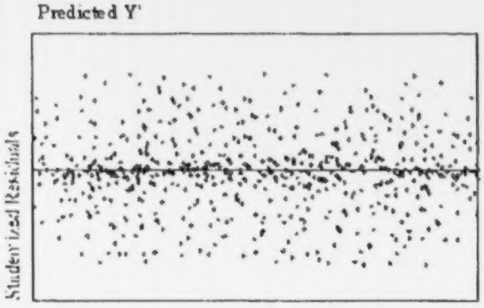
ASSUMPTION OF HOMOSCEDASTICITY

Homoscedasticity means that the variance of errors is the same across all levels of the IV. When the variance of errors dif different values of the IV, heteroscedasticity is indicated. According to Berry and Feldman (1985) and Tabachnick and Fi (1996) slight heteroscedasticity has little effect on significance tests; however, when heteroscedasticity is marked it can le serious distortion of findings and seriously weaken the analysis thus increasing the possibility of a Type I error.

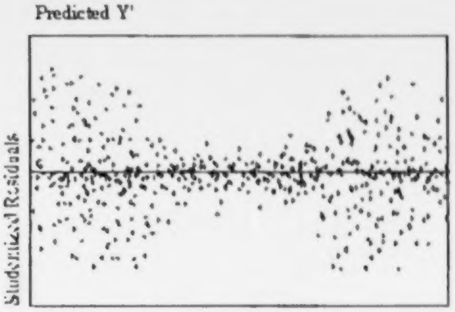
This assumption can be checked by visual examination of a plot of the standardized residuals (the errors) by the regressior standardized predicted value. Most modern statistical packages include this as an option. Figure 3 show examples of plot might result from homoscedastic and heteroscedastic data.

Figure 3. Examples of homoscedasticity and heteroscedasticity

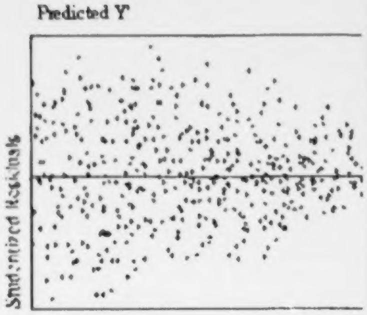
Homoscedasticity



Heteroscedasticity



Heteroscedasticity



Ideally, residuals are randomly scattered around 0 (the horizontal line) providing a relatively even distribution. Heterosce is indicated when the residuals are not evenly scattered around the line. There are many forms heteroscedasticity can take a bow-tie or fan shape. When the plot of residuals appears to deviate substantially from normal, more formal tests for heteroscedasticity should be performed. Possible tests for this are the Goldfeld-Quandt test when the error term either dec increases consistently as the value of the DV increases as shown in the fan shaped plot or the Glejser tests for heteroscedas when the error term has small variances at central observations and larger variance at the extremes of the observations as i bowtie shaped plot (Berry & Feldman, 1985). In cases where skew is present in the IVs, transformation of variables can r heteroscedasticity.

CONCLUSION

The goal of this article was to raise awareness of the importance of checking assumptions in simple and multiple regressio focused on four assumptions that were not highly robust to violations, or easily dealt with through design of the study, tha researchers could easily check and deal with, and that, in our opinion, appear to carry substantial benefits.

We believe that checking these assumptions carry significant benefits for the researcher. Making sure an analysis meets th associated assumptions helps avoid Type I and II errors. Attending to issues such as attenuation due to low reliability, curvilinearity, and non-normality often boosts effect sizes, usually a desirable outcome.

Finally, there are many non-parametric statistical techniques available to researchers when the assumptions of a parametri statistical technique is not met. Although these often are somewhat lower in power than parametric techniques, they provi valuable alternatives, and researchers should be familiar with them.

References

Berry, W. D., & Feldman, S. (1985). *Multiple Regression in Practice.* Sage University Paper Series on Quantitative App in the Social Sciences, series no. 07-050). Newbury Park, CA: Sage.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, N Lawrence Erlbaum Associates, Inc.

Nunnally, J. C. (1978). *Psychometric Theory* (2nd ed.). New York: McGraw Hill.

Osborne, J. W., Christensen, W. R., & Gunter, J. (April, 2001). Educational Psychology from a Statistician's Perspective: Review of the Power and Goodness of Educational Psychology Research. Paper presented at the national meeting of the

Education Research Association (AERA), Seattle, WA.

Osborne, J. W. (2001). A new look at outliers and fringeliers: Their effects on statistic accuracy and Type I and Type II er
Unpublished manuscript, Department of Educational Research and Leadership and Counselor Education, North Carolina S
University.

Pedhazur, E. J., (1997). *Multiple Regression in Behavioral Research* (3rd ed.). Orlando, FL:Harcourt Brace.

Tabachnick, B. G., Fidell, L. S. (1996). *Using Multivariate Statistics* (3rd ed.). New York: Harper Collins College Publis

Tabachnick, B. G., Fidell, L. S. (2001). *Using Multivariate Statistics* (4th ed.). Needham Heights, MA: Allyn and Bacon

Contact Information:

Jason W. Osborne, Ph.D
ERLCE, Campus Box 7801
Poe Hall 608,
North Carolina State University
Raleigh NC 27695-7801

(919) 515-1714

Jason_Osborne@ncsu.edu

Descriptors: Hypothesis Testing; *Regression [Statistics]; Research Methodology; Statistical Studies

```
ADODB.Recordset error '800a0e78'

Operation is not allowed when the object is closed.

/pare/getvn.asp, line 179
```

Home | Articles | Subscribe | Review | Policies

Volume: 8 7 6 5 4 3 2 1

Practical Assessment,
Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ☐ Go

Copyright 2002, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Brem, Sarah (2002). Analyzing online discussions: ethics, data, and interpretation. *Practical Assessment, Research & Evaluation*, 8(3). Available online: http://ericae.net/pare/getvn.asp?v=8&n=3. This paper has been viewed 3288 times since 5/15/02.

▸ Find similar papers in
ERICAE Full Text Library
Pract Assess, Res & Eval
ERIC RIE & CIJE 1990-
ERIC On-Demand Docs

▸ Find articles in ERIC written by
Brem, Sarah

Analyzing Online Discussions: Ethics, Data, and Interpretation

Sarah K. Brem
Division of Psychology in Education
Arizona State University

Online discussions are attractive sources of information for many reasons. Discussion forums frequently offer automated tracking services, such as a transcript or an archive, so that you can engage in animated conversation and analyze it at leisure, or locate conversations that took place months or years ago. Online tools provide an opportunity to observe a group without introducing your own agenda, to follow the development of an issue, or to review a public exchange that took place in the past, or outside the influence of researchers and policymakers. You can test additions and revisions to tools for communication, building more effective online classrooms, research groups, and professional organizations. Whether you are looking for ways to improve interactions within a working group (A[illegible] & Carley, 1998), studying the interactions of a community that interests you (Klinger, 2000), or assessing student learning (Brem, Russell, Weems, 2001), online discussions can be a valuable tool.

An online discussion is identified by the use of a computer-mediated conversational environment. It may be synchronous, such as real-time chat, or instant messaging, or asynchronous, such as a listserver, or bulletin board. It may be text-only, or provide facilities for displaying images, animations, hyperlinks, and other multimedia. It may require a Web browser, a Unix connection, or special software that supports such features as instant messaging. Tools for online conversation are becoming increasingly sophisticated, popular, and available, and this increases the appeal of using online discourse as a source of data.

Online discussions present new opportunities to teachers, policymakers, and researchers, but they also present new concerns and considerations. This article is about access to, and management and interpretation of, online data. Online research is similar, but not identical to, face-to-face (f2f) research. There are new ethical considerations that arise when it is not clear whether the participants in a conversation know they are being monitored, or when that

BEST COPY AVAILABLE

monitoring is so unobtrusive that it can easily be forgotten. Instead of collecting data using audio and video recording as in f2f conversations, preserving online conversations requires ways to download or track the electronic files in which the information is stored. Finally, in f2f interactions we examine body language and intonation as well as the words spoken, and in an online interaction, we have to look beyond the words written to the electronic equivalents of gestures and social conventions. This article will address these issues of ethics, data collection, and data interpretation.

This article is *not* about recommending any particular method of analysis. Whether you use grounded theory, quantifying techniques, experimental manipulations, ethnography, or any other method, you will have to deal with issues of collecting and managing data, as well as the structure of online communication. (For information about analyzing discourse, see Stemler, 2001; techniques and considerations that are specific to online discourse can be found in the 1997 special issue of *The Journal of Computer Mediated Communication*, "Studying the Net."). Information about tools for theory-based data manipulation is available at http://kerlins.net/bobbi/research/qualresearch/researchware.html and http://directory.google.com/Top/Science/Social_Sciences/Methodology/Qualitative/Tools/.

Ethical Considerations

Before we consider how to analyze an online conversation, we need to first consider what precautions should be taken to protect participants in the conversation. Because online conversation is relatively new and unfamiliar, and takes place at a distance, it is relatively easy to overlook possible ethical violations. People may not realize that their conversations could be made public, may not realize that they are being monitored, or may forget that they are being monitored because the observer's presence is virtual and unobtrusive. Some participants may feel relatively invulnerable because of the distance and relative anonymity of online exchanges, and may use these protections to harass other participants. Online exchanges of information require the same levels of protection as f2f exchanges, but it can be more complicated to achieve this.

If you belong to a university or similar institution, you will need the approval of an Institutional Review Board, created for the protection of human beings who participate in studies. Teacher-researchers and others who do not have an IRB and are not associated with any such institution should nevertheless follow the ethical principles and guidelines laid out in *The Belmont Report*, available at http://ohsr.od.nih.gov/mpa/belmont.php3. Other useful resources include Sales and Folkman (2000), and NIH ethics resources at http://www.nih.gov/sigs/bioethics/researchethics.html.

The least problematic conversations are those that take place entirely in the public domain; people know they are publishing to a public area with unrestricted viewing, as if they were writing a letter to the editor. Newsgroups are an example of such exchanges—anyone with access to http://groups.google.com/ can access any conversation in the past twenty years. In many cases, this sort of research is considered "exempt" under Federal guidelines for the protection of human subjects; for researchers at institutions with an IRB, the board must confirm this status. Still, even public areas may contain sensitive information that the user inadvertently provided; novices are especially prone to accidentally giving out personal information, or including personal information without considering possible misuse. In addition to the usual procedures for anonymizing data (e.g., removing names, addresses, etc.), there are some additional concerns to address. Every post must be scoured for both intentional and unintentional indicators of identity. Here are some common ways that anonymity is compromised:

- Usernames like "tiger1000" do not provide anonymity; people who are active online are as well known by their usernames as their traditional names. Usernames must be replaced with identifiers that provide no link to the actual participant.
- You must also be vigilant in removing a participant's .sig (the signature file that is appended to a post) and

any other quotes, graphics, and other idiosyncratic inclusions that are readily identifiable as belonging to a particular individual.

- Identifying information is often embedded in a post through quoting; for example, if I were quoted by another participant, my email address might be embedded in the middle of his or her message as "tiger1000 (sarah.brem@asu.edu) posted on 1 February 2002, 11:15."

If a domain establishes any degree of privacy through membership, registration, passwords, etc., or if you wish to contact participants directly, then the communications should be considered privileged to some degree. In addition to the safeguards required for public domain data, using these conversations in research requires at very least the informed consent of all participants whose work will be included in the analysis, with explicit description of how confidentiality and/or anonymity will be ensured. The procedures for informed consent, recruitment, and data collection will require "expedited" or "full" review by an Institutional Review Board. Once approval has been given, consent forms will have to be distributed to every participant, and only the contributions of consenting members can be stored and analyzed.

If you set up a site for collecting data, regardless of how much privacy and anonymity you promise, you are ethically bound to inform all potential participants that their contributions will be used as data in research. One example of how to provide this information has been implemented by the Public Knowledge Project. To see how they obtained consent, visit http://www.pkp.ubc.ca/bctf/terms.html. Likewise, if you contact participants directly, you need to make their rights clear and obtain their permission to use the information they provide for research purposes before engaging in any conversation with them.

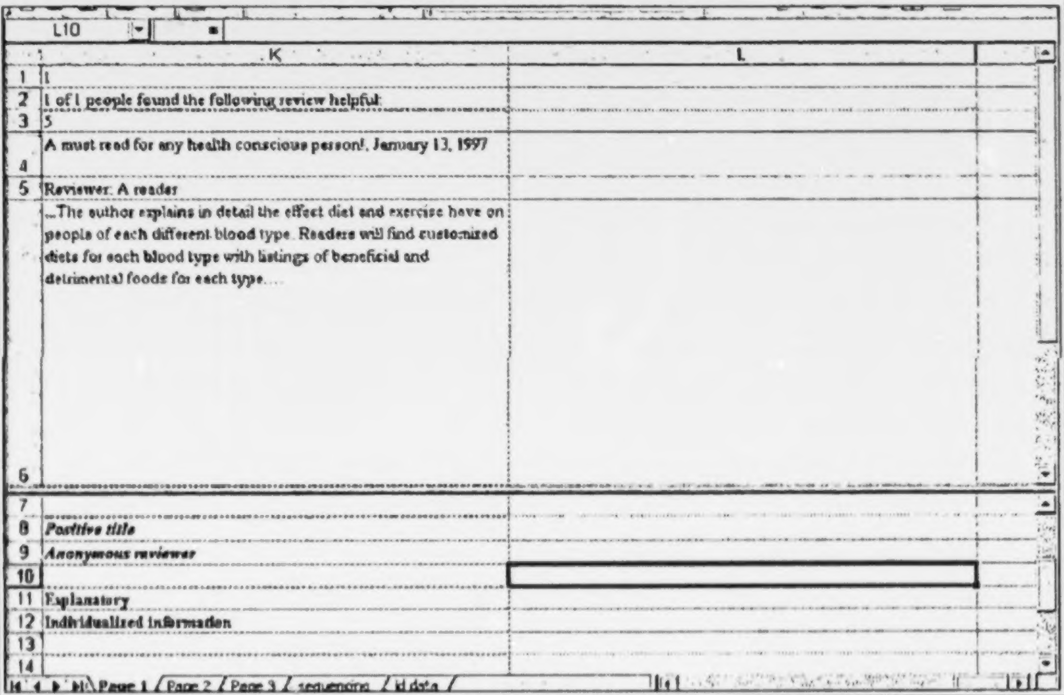
In addition to preserving the safety and comfort of participants, you must also consider their intellectual property rights. All postings are automatically copyrighted under U.S. and international laws. Extended quotes could violate copyright laws, so quoting should be limited, or permission should be obtained from the author prior to publication. For more about U.S. and international laws, visit http://www.law.cornell.edu/topics/copyright.html.

Data Collection and Management

Once you have received the necessary permissions and taken the necessary precautions, the next concern is the best way to collect and organize the data for analysis. An online exchange often evolves over days or months, and may require handling tens of thousands of lines of text, along with graphics, hyperlinks, video, and other multimedia. Consider what media will be present before choosing tools for management and manipulation.

For text-only exchanges, a flatfile spreadsheet is often sufficient. The text can be downloaded as plaintext, or cut and pasted in sections. Paragraphs or lines of text become entries in the spreadsheet, and can be parsed into smaller units if desired. Once the data is placed in a spreadsheet, additional rows and columns can be used to hold codes and comments, and the spreadsheet can be sorted on these to reveal and examine patterns. An example of how this can be done is presented in Figure 1.

BEST COPY AVAILABLE

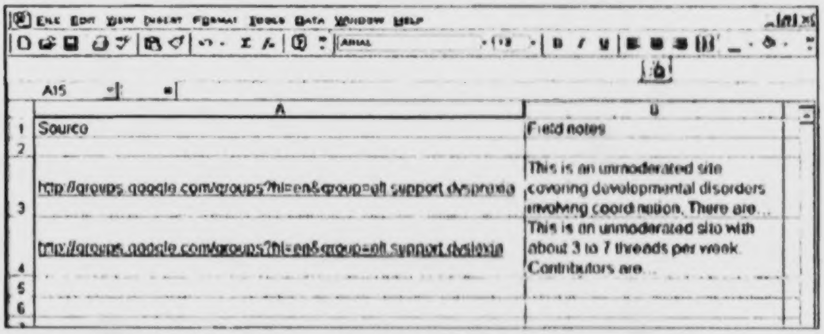


L10

	K	L
1	1	
2	1 of 1 people found the following review helpful:	
3	5	
4	A must read for any health conscious person!, January 13, 1997	
5	Reviewer: A reader	
6	...The author explains in detail the effect diet and exercise have on people of each different blood type. Readers will find customized diets for each blood type with listings of beneficial and detrimental foods for each type....	
7		
8	*Positive title*	
9	*Anonymous reviewer*	
10		
11	Explanatory	
12	Individualized information	
13		
14		

Figure 1. This sample is taken from an analysis of an online discussion of a book. At the top of the screen shot is a participant's entry, along with the information regarding its position in the thread, date of posting, and so on. Codes were added below each entry, shown at the bottom of the screen shot. The color of the codes corresponds to the color of the relevant text. Each participant's contribution can be added as an additional column.

There are many cases, however, when this technique will be ineffective. Because they can last for years, online conversations differ from f2f conversations in that they can be extremely long, often exceeding spreadsheet limits. Furthermore, they often contain hyperlinks, graphics, video, and other multimedia; these are often essential to the conversation, and most spreadsheets will not display them. When it is desirable to maintain these elements, there are two straightforward ways to do this. The first is to simply download all the relevant files and create a mirror archive on your own hard drive. This assures you constant, reliable access to the data, but may take up large amounts of space, and not all files can be downloaded (e.g. there may be security restrictions, additional software requirements, or intellectual property considerations). An alternative approach is to create a flatfile spreadsheet that contains hyperlinks to the original exchanges rather than the exchanges themselves. The disadvantage is that you cannot be sure the original files will always be available, but the spreadsheets containing these pointers take up very little space, are less problematic legally and technologically, and provide the full functionality of a spreadsheet (e.g., sorting and manipulation).



	A	B
1	Source	Field notes
2		
3	http://groups.google.com/groups?hl=en&group=alt.support.dyspraxia	This is an unmoderated site covering developmental disorders involving coordination. There are...
4	http://groups.google.com/groups?hl=en&group=alt.support.dyslexia	This is an unmoderated site with about 3 to 7 threads per week. Contributors are...
5		
6		

Figure 2. This example includes links to discussions on developmental disabilities that affect schoolchildren. The links take you to the conversation described in the field notes.

The advantage to using a flatfile database is that it allows for flexible coding. The disadvantage is that it does not support any particular theoretical perspective. For this reason, you may want to begin by using a flatfile, then transfer data to a theory-based format after you have done some initial processing and can narrow down what you want to focus on. Such tools are described at the sites mentioned above.

Data Preparation, Manipulation, and Preservation

Online data creates extremely large files, not only because of their potential length but also because online conversations tend to be highly repetitive. Replies often contain portions of previous messages, if not the complete original; even if each individual contribution is relatively short, quoting can quickly create messages containing hundreds of lines. In addition, multimedia elements tend to take up considerable space. It is not unusual for a datafile to grow to 30 megabytes or more. Files of this size are very difficult to manipulate and can be prohibitively slow to load and save. Therefore, it may become necessary to decide what information should be kept verbatim, what should be deleted altogether, and what can be replaced with a smaller reference code (e.g., if many participants quote message 112, you might replace each reposting of this message with the code "msg112"; advertisements might be indicated by the code "banner ad" or a hyperlink to the ad on the original site). These methods of abridging the record can be implemented before engaging in extensive analysis, so that the file that you work with most often is the smallest one.

In deciding on these matters, you should be guided by your research questions and you should preserve all information that is relevant to your questions; thus, advertising may be a central issue, or it may play a relatively small role. In any case, it is best to err on the side of preserving too much information. Once removed, a hyperlink, graphic, or reposted message can be difficult to recover. Start by keeping as much information as possible, and pare it down as you find elements that seriously interfere with speed, or that are adding nothing to your analysis. You may want to keep multiple versions, using the most streamlined for your analysis, and archiving richer versions in case they are needed later on.

Coding, Analysis, and Interpretation

The structure of an online exchange can be difficult to reconstruct, and its boundaries can be difficult to locate. Capturing the perspective of participants, challenging in any context or medium, is further complicated by new ambiguities created by the way in which conversations are created, stored, and accessed. While it may not be possible to resolve all inconsistencies and ambiguities, being aware of them and their implications for any particular interpretation is essential.

BEST COPY AVAILABLE

Reconstructing the Conversation

One significant difference between online and f2f conversations is that participants often view online conversations differently. Online discussions do not necessarily develop sequentially, nor can we be sure that all participants are seeing the same exchange. We can see this by comparing how listservs and bulletin boards are visited and revisited. A listserv sends messages to the subscriber's email account. Listservs send all messages in chronological order, regardless of the conversational thread to which they belong, so multiple conversations are interleaved. It is easy to miss a post, and each person may read a different set of messages. If you join a listserv after a conversation has begun, you will not see the beginning of the exchange. In contrast, bulletin boards keep message separate by thread, and all messages are available for the life of the bulletin board, or until they are archived.

A participant may follow conversations thread by thread, read everything written by a single author, skip from one topic to the next, or otherwise deviate from the original presentation. You should consider reviewing the conversation in a variety of ways in order to understand better how participants receive and work with the information.

For example, Usenet groups often attract users who only wish to ask a single question, get an answer, and never return. In addition, while some servers provide a complete, searchable Usenet archive (http://groups.google.com/), others regularly delete files to save space, or may not provide much in the way of searchability. For these reasons, it is common for several participants to ask the same question, sometimes word for word, over and over. Understanding why this happens and how the conversation develops requires looking at the records both as if you are a user with access to the full record and, as if you are a user with access to a very limited record. It is virtually impossible to capture all possible viewings, but you will probably want to capture several.

Tracking a conversation, regardless of the perspective you choose, can be challenging, rather like assembling a rough-cut jigsaw puzzle. The threads of conversation are easily broken; if a participant or server changes a subject line, archiving tools cannot follow the conversation and the line of thought becomes disconnected. People use multiple accounts and identities, either because they are deliberately trying to hide their identity, or for innocent reasons, such as logging in differently from work and home. There are, however, ways to reconstruct a conversation. To track a thread, examine subject lines to see if they correspond except for a reply indicator, look at dates of posting, or examine the text for quotes from previous messages in the thread or other references to previous postings in the thread. In the case of users, even if participants' usernames change, they may be identifiable through their email addresses, their signatures, hyperlinks to their home pages, or their writing styles and preferred themes. For example, in analyzing one Usenet group in which the topic of speed reading frequently arose, I noted that there were several usernames employed by one company; these users would respond as if they were "ordinary" individuals, rather than identifying themselves as company representatives. However, all used the same prefabricated plug for the company's product. Thus, I could use this to mark the posts as coming from related users, or perhaps the same user.

Where, What, and Who is the Conversation?

In addition, consider the context. F2f conversations consist of a relatively well-bounded exchange; the participants, location, and duration are easier to determine than they are in online discourse. Online, participants can possess multiple identities, steal another's identity, or carry on a conversation with themselves. The conversation not only crosses geographical boundaries, but may send participants to archives of prior exchanges, websites, FAQs, and other resources. As a result, the conversation may not be neatly contained by a single listserv, chat room, or other discourse environment. Even within a single environment, the conversation you are interested in may be no more than a few lines or posts tucked in among other threads, spam (mass mailings), flames (inflammatory posts), and announcements. Finally, regarding duration, online conversations may last minutes or years, and may proceed at the rate of several exchanges per minute or one exchange every few weeks or months.

Given these complexities, the best approach is to be aware that you will have to draw somewhat arbitrary boundaries around a group of participants and exchanges, letting your choice be led by your questions. If identifying participants is crucial (perhaps you suspect that warring factions are trying to discredit one another by posing as members of the other camp), then you will have to look for clues that reveal identity and consider how your interpretations are affected by the possibility of imposters. If the conversation takes place amongst a small, tightly knit group with a strong foundation of common knowledge, then shared spaces like FAQs and group websites becomes crucial, and should be included. If there have been significant changes in the political or educational climate during the course of the conversation, duration will become important, and the timeline of the exchange may need careful examination.

You will always have to draw boundaries, and there will never be one right set of boundaries to draw. The important thing is to draw them in such a way that you can explain your reasoning to others, and in a way that allows you to get rich, useful, and dependable answers to the questions that interest you.

Knowing How to Talk Online

We do not analyze f2f conversations without having some experience with f2f conversation, both at an everyday level and at the more finely honed level of a discourse expert. You should also become a participant in online communities before trying to research them, gaining both everyday and scholarly familiarity. Rather than just knowing the basics of navigation and communication, it is important to be fluent in everyday conventions and the online analogs of body language and nonverbal f2f communication. These include "emoticons" (e.g., symbols for smiling `8^)`, disgust `8-P`, and so on), as well as conventions such as SCREAMING BY TYPING IN ALL CAPS, or including actions, such as ::hugs newbie:: or <<grins at newbie>>. (For more information about communication conventions on the Internet, visit http://www.udel.edu/interlit/chapter5.html.)

In addition, learn how to relate to participants as individuals; it is easy to fall into the trap of treating them as disembodied voices, or automatons, rather than as complete people. What are their interests online and off? Is their style of conversation friendly, combative, joking, pedantic? What topics will get an emotional reaction from them? What sorts of conversational moves will get a reaction from them (e.g., some people are offended by posts IN ALL CAPS, and will tell the poster to stop shouting)? In an extended conversation with a group, you should get to a point that you can recognize participants without relying solely on usernames.

Final Thoughts

The study of online discourse is still quite new, and there is much about the treatment and analysis of these data that has not yet been addressed. When faced with a situation for which there is no standard procedure, the best course of action is to begin with established techniques and then adapt these to the online environment. Have a rationale for any adaptations or deviations you decide to make because these will help you to establish credibility with editors and peers and will allow others to adopt, recycle, and refine your approach.

Notes:

This work was supported by an NSF Early Career Award (REC-0133446). Investigating Critical Thinking In Multimedia Environments To Improve Public Utilization Of Science.

References

Ahuja, M. K. & Carley, K. M. (1998). Network structure in virtual organizations. *Journal of Computer-Mediated*

Communication, 3 (4). Available online: http://www.ascusc.org/jcmc/vol3/issue4/ahuja.html.

Brem, S. K., Russell, J. & Weems, L. (2001). Science on the Web: Student evaluations of scientific arguments. *Discourse Processes*, 32, 191-213. Available online: http://www.public.asu.edu/~sbrem/docs/BremRussellWeems.webversion.htm.

Klinger, S. (2000). "Are they talking yet?": Online discourse as political action. *Paper presented at the Participatory Design Conference*, CUNY, New York. Available online: http://www.pkp.ubc.ca/publications/talkingyet.doc.

Sales, B. D. & Folkman, S. (2000). *Ethics in research with human participants*. Washington, DC: American Psychological Association

Stemler, Steve (2001). An overview of content analysis. *Practical Assessment, Research & Evaluation*, 7(17). Available online: http://ericae.net/pare/getvn.asp?v=7&n=17.

Descriptors: Content Analysis; Research Methods; World Wide Web

```
ADODB.Recordset error '800a0e78'

Operation is not allowed when the object is closed.

/pare/getvn.asp, line 179
```

Home | Articles | Subscribe | Review | Policies

Volume: 8 7 6 5 4 3 2 1

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ☐ Go

Copyright 2002, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Kester, Ellen Stubbe & Elizabeth D. Peña (2002). Language ability assessment of spanish-english bilinguals: future directions. *Practical Assessment, Research & Evaluation*, 8(4). Available online: http://ericae.net/pare/getvn.asp?v=8&n=4. This paper has been viewed 4546 times since 5/18/02.

▸ Find similar papers in
ERICAE Full Text Library
Pract Assess, Res & Eval
ERIC RIE & CIJE 1990-
ERIC On-Demand Docs

▸ Find articles in ERIC written by
Kester, Ellen Stubbe
Elizabeth D. Peña

Language ability assessment of Spanish-English bilinguals: Future directions

Ellen Stubbe Kester and *Elizabeth D. Peña*
The University of Texas at Austin

Children from non-English speaking backgrounds are often misdiagnosed with language impairment due to a number of reasons. One of the primary reasons is that currently, there are limited diagnostic tools available that are designed for children who are exposed to two languages (Valdés & Figueroa, 1994). Current practices for assessment of language in bilinguals frequently involve the use of tests that are translated from English to the target language and/or tests designed for and normed on monolinguals. These currently available tools are not well suited for a bilingual population because they do not take into account the unique aspects of bilingual language acquisition. While the focus of this paper is on language assessment of bilinguals for the purpose of differentiating language impairment from typical language development, the issues presented have implications for all fields that include language as part of the assessment process, including IQ, educational, and achievement testing.

The objectives of this paper are to a) summarize relevant research on bilingual language development and discuss the implications for bilingual language assessment, b) discuss limitations in current language ability testing practices for bilinguals, c) propose future directions for the development of assessment tools and practices with bilinguals.

Research on Bilingual Language Development: Implications for Assessment

Generally, the testing practices used today for bilinguals operate under the assumption that there is no difference in the language development of monolinguals and bilinguals. However, research in the area of bilingual language development suggests that bilinguals have different patterns of language development than monolinguals of either language (Grosjean, 1989). Consistent with the *Competition Model* (CM), Hernandez, Bates and Avila (1994) proposed that bilinguals use an amalgamation of strategies used by monolingual speakers. A key component of the

BEST COPY AVAILABLE

CM model is that there are competing cues in any given language that help map meaning to utterances. The informational value of cues is determined by the frequency with which this type of information is available during decision-making processes and the frequency with which this type of information leads to a correct conclusion when it is used. As applied to language development, children test the use of different cues before they establish which cues best yield interpretations that are consistent with their environment.

Cross-linguistic studies based on this model indicate that the cues used to process and produce language efficiently are not the same across languages (MacWhinney & Bates, 1989). Examples of cues include word order and subject-verb agreement. In English, word order is relatively strict as compared to Spanish. The following sentence, "The boy (subject) hit (verb) the ball (object)" has a different meaning from "The ball (subject) hit (verb) the boy (object), because of word order cues. An English speaker would identify the subject (boy or ball) as the one hitting due to its position in the sentence. Spanish speakers, on the other hand, rely less on word order cues. For example, *El niño* (subject) *comió* (verb) *los frijoles* (object) (The boy ate the beans) has the same meaning as *Comió* (verb) *los frijoles* (object) *el niño* (subject). In comparison to English, Spanish has a complex verb system in which the verb stem provides cues about the subject, tense, and mood of the sentence. However, English verb morphology provides fewer cues about the subject. For example, the verbs *comí, comiste, comió,* and *comieron* are all represented by the verb *ate* (I ate, you ate, he ate, they ate, respectively) in English. Thus, bilingual children must learn how cues work within and between their two languages, creating a unique system of cues drawn from two languages (i.e., an amalgamated system). When children are developing two languages they often apply cues from L1 to L2 and from L2 to L1. Thus, bilingual children follow a different developmental course of language development in each of their languages in comparison to monolingual children. Language tests for bilinguals should reflect these differences in development.

Limitations of Current Language Testing Practices for Bilinguals

Two common practices in the language assessment of bilinguals are translations of tests and the use of tests designed for monolinguals of the child's native language and/or second language. However, evidence that different linguistic cues are prominent in different languages and that bilinguals likely use an amalgamated cue system, suggests that translated tests and tests normed on monolinguals are likely to yield invalid estimates of language ability in bilinguals.

Problems with Test Translation

When tests are translated from one language to another, they do not retain their psychometric properties. Of particular interest in the assessment of language is the developmental order in which target features of the language are learned. Translating a test from one language to another -- typically from English -- may mean that items are organized by order of English difficulty, rather than reflecting the developmental order of the target language. The translated Spanish version of the Preschool Language Scale-3 (Zimmerman, Steiner, & Pond, 1993) provides an illustration. Restrepo and Silverman (2001) found several item difficulty discrepancies between the original English and the translated Spanish version when tested with predominately Spanish-speaking preschoolers. For example, items related to prepositions, which were relatively easy for English speakers, were more difficult for Spanish speakers. On the other hand, the "function" items were easier for the Spanish speakers in comparison to the English speakers.

The notion of cue validity can be used to examine development of semantic representation. Figueroa (1989) noted that words may generally represent the same concept but have variations and different levels of difficulty across languages, possibly due to their prominence, information load, and/or frequency. An illustration of this is found in a study of vocabulary test translations (Tamayo, 1987). When test items were translated from English to Spanish they differed in frequency of occurrence in each language. Because the Spanish translations were of lower frequency within Spanish, test scores obtained from Spanish speakers were lower compared to scores obtained from the original English version. However, when the vocabulary items were matched for their frequency of occurrence in the original and target language and matched for meaning, test scores obtained from Spanish and English speakers were equivalent.

Similarly, the context in which words are learned influences category development. Across different languages, the same general category may have different prototypical members, and different words may be associated with each language for the same situation. These contextual variations make translated vocabulary tests particularly vulnerable to imbalance. In a category generation task with bilingual four to six year-olds, Peña, Bedore, and Zlatic-Giunta (in press) found that for animals, children's three most frequent English responses were "elephant," "lion," and "dog," while in Spanish they used "caballo" (horse), "elefante" (elephant), and "tigre" (tiger) in these orders. Clearly, the circumstances under which children learn language affect their representation of language.

In addition to vocabulary differences, grammatical structure also affects the validity of test translation practices. For example, nouns are marked by gender in Spanish but not English, resulting in different cue values for each language. An English test translated to Spanish will miss aspects of Spanish, such as gender marking, that are not present in the English language. Furthermore, in Spanish, subject information is frequently carried in the verb, resulting in more complex verbs and less salient pronouns as compared to English. In English language assessment, pronoun omission is a hallmark of language impairment, yet this would not be true for Spanish. Thus, translated language tests may target inappropriate features for the target language, resulting in inaccurate assessment of language ability.

Problems comparing bilinguals and monolinguals

Bilingual school children generally fall into the category of circumstantial bilinguals. That is, their circumstances (often a Spanish-speaking home environment and an English-speaking or bilingual school environment) require them to use two languages. These different environments typically require different language content. The home environment likely promotes discussions of common family activities, such as cooking or trips to the store, while more academic topics, such as colors, numbers, and shapes, are highlighted in the school environment. As such, bilingual children will develop different vocabulary content for each language. From a testing perspective, this can result in underestimation of concept knowledge when testing in only one language at a time, or even when testing in both languages.

For example, Sattler and Altes (1984) examined typically developing three to six year-old bilingual Latino children's scores on the Peabody Picture Vocabulary Test-Revised and the McCarthy Perceptual Performance Scale. They found that the PPVT-R, whether administered in English or Spanish, yielded scores far below those of the norms, while all of the children were estimated to have normal intelligence based on their McCarthy scores.

Further investigation of the research on vocabulary development in bilinguals provides evidence of their use of a unique bilingual profile, and is consistent with the notion of an amalgamated rather than a "two monolinguals in one" system. A number of studies in the area of vocabulary acquisition illustrate that in early development, bilinguals learn unique words across their two languages, rather than learning two words (one in each language) for each concept. Pearson, Fernández, and Oller (1992) found that young bilinguals (8-30 months) often produced words for different concepts in each language, with few concepts labeled in both languages. Similarly Peña, Bedore, and Zlatic (in press) found that in a category generation task, bilingual children (ages 4-6 years) produced more unique words across Spanish and English (referred to as a *conceptual score*), in comparison to doublet (overlapped) words.

When monolinguals and bilinguals are compared on measures of vocabulary, differences become more apparent. Pearson, Fernández, and Oller (1993) used the Spanish and English versions of the MacArthur Communicative Development Inventory (1989) to estimate bilingual toddler's vocabularies. They found that when compared to monolingual norms in either language, their scores were low. However, when they compared the total number of unique words they produced across the two languages, their scores were more comparable to the monolingual norms.

Another example of findings of differential performance between monolinguals and bilinguals is with the Test de Vocabulario en Imágenes Peabody: Adaptación Hispanoamericana (TVIP-H; Dunn, Padilla, Lugo, & Dunn, 1986). This version of the Peabody Picture Vocabulary Test (PPVT; Dunn, 1959) was normed on monolingual Spanish speakers outside of the U.S. mainland and then tested with bilingual Hispanics on the U.S. mainland. Results were that the bilinguals' scores were lower than those of the monolinguals (Dunn, 1988). Over age, the differences between monolinguals and bilinguals increased and coincided with schooling in English. Similarly, Umbel, Pearson, Fernandez, and Oller (1992) used the Peabody Picture Vocabulary Test-Revised (Dunn & Dunn, 1981) in English and the complementary Spanish version, the Test de Vocabulario en Imágenes Peabody (TVIP-H), to compare the receptive vocabularies of bilingual children (ages 5 years 11 months to 8 years 6 months) who were exposed to both Spanish and English in the home. Findings were that children on average responded correctly to 67% of the items in their age range in both languages, but that another 8% to 12% were known only in one of their two languages. Administration of this test in only one language -- even the "dominant" language -- would have led to an underestimation of vocabulary knowledge.

Conceptual scoring (Pearson, Fernández, & Oller, 1993) has been proposed as a more meaningful measure of the bilingual's conceptual knowledge. The system, which entails counting the concepts demonstrated (either through constructed or selected responses) in both languages and correcting for concepts shared in the two languages, results in a more valid representation of a bilingual child's knowledge of concepts.

Future Directions

Item difficulty values, item discrimination, reliability, and validity are affected when tests are translated. For example, item difficulty values are affected when "equivalent" lexical items differ in frequency of occurrence (Tamayo, 1980). Less-frequent words have higher difficulty, while more frequent words are generally easier. Similar patterns of changes in item difficulty are seen for items that address conceptual framework, grammatical structure, and specific social content. The documented differences in bilingual and monolingual language development provide evidence suggesting that use of translated tests or tests designed for monolinguals will result in questionable validity. Clearly, the psychometric properties of a test do not translate from one language to another, nor do they remain the same when the test is administered to a different audience than intended.

While improving translation practices and uses of tests designed for monolinguals is an important short-term goal, long-term goals should include the development of language tests designed for, and normed on, bilinguals. In order to achieve such a goal, future research is needed to better understand the development of semantic and syntactic language skills in bilinguals. We offer the following recommendations to test developers:

- **Sample domains broadly during the exploratory level of test development to ensure that concepts and linguistic features are appropriately represented for each language.** For example, tests of semantic language skills should explore a wide variety of semantic concepts, such as similarities and differences in objects, functions of objects, categorization, characteristic properties, word associations, and spatial relations. Tests of grammar should explore a wide variety of structures in both languages rather than focusing on only the structures the two languages have in common, or on only structures important in English. Clinically, these suggestions apply as well. Testing beyond the ceiling, using dynamic assessment, clinical interviewing, and feedback during or as a follow-up to assessment of bilinguals may help better estimate true language ability.

- **Use conceptual scoring systems to eliminate underestimation of ability.** When testing concepts, consider a bilingual child's conceptual system as a whole, rather than as two language-specific systems. Thus, a bilingual approach accounting for the commonalities and differences across two languages is recommended over two monolingual assessments. When different concepts are expressed across languages, all should be counted. An example of an attempt at considering two languages is the English/Spanish Bilingual Verbal Ability Tests (BVAT) (Cummins, Muñoz-Sandoval, Alvarado, & Ruef, 1998), which assumes that bilinguals have a unique linguistic configuration, rather than two language-specific configurations. The BVAT estimates a bilingual's verbal ability by measuring the linguistic knowledge common to the bilingual's two languages and the linguistic knowledge unique to each language.

- **Select an appropriate mix of item types to gain the maximal amount of information about language ability in each language.** Rather than try to balance item types across languages, consider that some types of items may be more appropriate targets in one language than the other. For example, an English grammar test might include more items related to pronouns than a Spanish test because pronouns are more salient in English, whereas a Spanish grammar test might include more items related to gender and number agreement.
- **When trying to balance concepts in different language versions of tests, consider the frequency of occurrence of the words.** There are a number of published materials on word frequency in different contexts available in both Spanish and English that can be used to ensure that "equivalent" terms are not only equivalent in meaning but in frequency (or difficulty) as well.

References

Cummins, J., Muñoz-Sandoval, A.F., Alvarado, C.G., & M.L. Ruef (1998). *The Bilingual Verbal Ability Tests.* Itasca, IL: Riverside.

Dunn, L. H. (1959). *Peabody Picture Vocabulary Test.* Circle Pines, MN: American Guidance Services.

Dunn, L. H. (1988). *Bilingual Hispanic Children on the U. S. Mainland: A Review of Research on Their Cognitive, Linguistic, and Scholastic Development.* Honolulu, HI: Dunn Educational Services.

Dunn, L. M., & Dunn, L. M. (1981). *Peabody Picture Vocabulary Test—Revised.* Circle Pines, MN: American Guidance Services.

Dunn, L. M., Padilla, E. R., Lugo, D. E., & Dunn, L. M. (1986). *Test de Vocabulario en Imágenes Peabody: Adaptación Hispanoamericana.* Circle Pines, MN: American Guidance Service.

Figueroa, R. (1989). Psychological testing of linguistic-minority students: Knowledge gaps and regulations. *Exceptional Children, 56,* 145-148.

Grosjean, F. (1989). Neurolinguists, Beware! The Bilingual is Not Two Monolinguals in One Person. *Brain and Language, 36*, 3-15.

Hernandez, A. E., Bates, E., & Avila, L. X., (1994). On-line sentence interpretation in Spanish-English bilinguals: What does it mean to be "in between"? *Applied Psycholinguistics, 15,* 417-46.

MacArthur Communicative Development Inventory. (1989). San Diego: University of California, Center for Research in Language.

MacWhinney, B. & Bates, E. (Eds.). (1989). *The Crosslinguistic Study of Sentence Processing.* New York: Cambridge University Press.

Pearson, B. Z., Fernandez, M. C., & Oller, D. K., (1992). Measuring bilingual children's receptive vocabularies.

Child Development, 63, 1012-1221.

Pearson, B. Z., Fernandez, M. C., & Oller, D. K., (1993). Lexical development in bilingual infants and toddlers: Comparison to monolingual norms. *Language Learning, 43*, 93-120.

Peña, E. D., Bedore, L. M., & Zlatic-Giunta, R. (in press). Development of categorization in young bilingual children. *Journal of Speech, Language, and Hearing Research.*

Restrepo, M. A., & Silverman, S. W. (2001). Validity of the Spanish Preschool Language Scale-3 for use with bilingual children. *American Journal of Speech-Language Pathology, 10,* 382-393.

Sattler, J. M., & Altes, L. M. (1984). Performance of bilingual and monolingual Hispanic children on the Peabody Picture Vocabulary Test—Revised and the McCarthy Perceptual Performance Scale. *Psychology in the Schools, 21,* 313-316.

Tamayo, J. (1987). Frequency of use as a measure of word difficulty in bilingual vocabulary test construction and translation. *Educational and Psychological Measurement, 47*, 893-902.

Umbel, V. M., Pearson, B. Z., Fernández, M. C., & Oller, D. K. (1992). Measuring bilingual children's receptive vocabularies. *Child Development, 63,* 1012-1020.

Valdés, G., & Figueroa, R. A. (1994). *Bilingualism and Testing: A Special Case of Bias.* Norwood, NJ: Ablex.

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (1993). *Preschool Language Scale-3: Spanish Edition.* San Antonio, TX: Psychological.

Send editorial correspondence to:

Ellen Stubbe Kester
Department of Communication Sciences and Disorders
Jesse H. Jones Communication Center, CMA 7.214
The University of Texas at Austin
Austin, TX 78712-1089

e-mail: stubbe.kester@mail.utexas.edu

Descriptors: Bilingualism; * Hispanic Americans; * Student Evaluation; * Second Languages; Bilingualism; * Gifted; * Hispanic Americans; * Screening Tests; * Second Languages

ADODB.Recordset error '800a0e78'

Operation is not allowed when the object is closed.

/pare/getvn.asp, line 179

Home | Articles | Subscribe | Review | Policies

Volume: 8 7 6 5 4 3 2 1

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ☐ Go

Copyright 2002, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Athanasou, James A. & Olabisi Olasehinde (2002). Male and female differences in self-report cheating. *Practical Assessment, Research & Evaluation*, 8(5). Available online: http://ericae.net/pare/getvn.asp?v=8&n=5. This paper has been viewed 5093 times since 5/18/02.

▸ Find similar papers in
ERICAE Full Text Library
Pract Assess, Res & Eval
ERIC RIE & CIJE 1990-
ERIC On-Demand Docs

▸ Find articles in ERIC written by
Athanasou, James A.
Olabisi Olasehinde

Male and female differences in self-report cheating

James A Athanasou
University of Technology, Sydney

Olabisi Olasehinde
University of Ilorin – Nigeria

Cheating is an important area for educational research, not only because it reduces the consequential validity of assessment results, but also because it is anathema to widely held public principles of equity and truthfulness (see Cizek, 1999 for a comprehensive review of the topic). Moreover, modern education is centred on numerous situations that really depend upon a student's honesty. The purpose of this paper is to review the extent of academic cheating and to describe any gender differences in self-reports.

There is a large literature relating to the influence of gender on academic cheating or dishonesty; the topic has captured the attention of researchers since the pioneering work of Hartshorne and May on deceit and honesty in 1928. It has been part of a wider fascination with the ethical, moral, and social dimensions of this academic behaviour at all levels of education.

In the course of various studies of academic honesty, it has been suggested that there are gender differences in the extent of cheating in education and that overall, women are less likely to cheat, but this conclusion has been challenged (Black, 1962; Graham, Monday, O'Brien & Steffen, 1994; Hartshorne & May, 1928; Kerkvliett, 1994; McCabe & Trevino, 1996). Good, Nichols, and Sabers (1999) noted that there may be differences in the perception of cheating between males and females, yet a meta-analytic study of gender and sex roles in relation to cheating produced a low mean effect size of 0.19 for self-reports of cheating (Whitley, Nelson & Jones, 1999). This paper considers some key issues in past research on cheating and reviews those studies that have investigated the influence

BEST COPY AVAILABLE

of gender on self-reported cheating.

Academic cheating

There is no consensus in estimates of the extent of cheating but it has been viewed as a major problem, with the majority of students indicating that they have been dishonest (Baird, 1980; Sierles, Hendrickx & Circle, 1980; Whitley, 1998). When specific forms of cheating such as plagiarism, collusion, copying, etc., are investigated, then the proportions reporting that they have been dishonest are reduced (see Hollinger & Lanza-Kaduce, p. 293). Kerkvliett and Sigmund (1999) discussed the prevalence of cheating within university and college systems and concluded: "The evidence indicates that many students cheat regularly and few students never cheat" (p. 331).

At the outset, it may be helpful to describe cheating for the reader as it includes a variety of behaviours. The essence of cheating is fraud and deception. We have adapted a working description of cheating in education contexts from a discussion on academic dishonesty in nursing students (Gabertson, 1997), as involving conscious participation in deception (through lying, dishonesty, falsifying, misrepresenting, corruption, plagiarism, copying, or unlawfully assisting someone else). Newstead, Franklyn-Stokes and Armstead (1996) provided the following list of cheating behaviours (see also Baird, 1980; Franklyn-Stokes and Newstead, 1995, p. 164), which we have classified according to Cizek (1999, p. 39):

Cheating by taking, giving, or receiving information from others

- Allowing own coursework to be copied by another student
- Copying another student's coursework with their knowledge
- Submitting a piece of coursework as an individual piece of work when it has actually been written jointly with another student
- Doing another student's coursework for them
- Copying from a neighbour during an examination without them realising
- Copying another student's coursework without their knowledge
- Submitting coursework from an outside source (e.g., a former student offers to sell pre-prepared essays, "essay banks")
- Premeditated collusion between two or more students to communicate answers to each other during an examination
- Obtaining test information from other students

Cheating through the use of forbidden materials or information

- Paraphrasing material from another source without acknowledging the original author
- Inventing data (i.e., nonexistent results)
- Fabricating references or a bibliography
- Copying material for coursework from a book or other publication without acknowledging the source
- Altering data (e.g., adjusting data to obtain a significant result)
- Taking unauthorised material into an examination

Cheating by circumventing the process of assessment

- Taking an examination for someone else or having someone else take an examination
- Attempting to obtain special consideration by offering or receiving favours through bribery, seduction, corruption
- Lying about medical or other circumstances to get special consideration by examiners (e.g., to get a more lenient view of results; extra time to complete the exam; an extended deadline; or exemption)

- Deliberately mis-shelving books or journal articles in the library so that other students cannot find them or by cutting out the relevant article or chapter
- Coming to an agreement with another student to mark each other's work more generously than it merits
- Illicitly gaining advance information about the contents of an examination paper
- Concealing teacher or professor errors
- Threats or blackmail or extortion

Consequently, cheating involves a wide range of behaviours. They can vary in their seriousness, execution, purpose, and social dimensions.

Early descriptions of honesty emphasised that it was situation-specific (Hartshorne & May, 1928). Later investigations of cheating behaviours have looked at relationships with factors such as culture, socialisation, field of study, extent of competitiveness and gender (Bowers, 1964; McCabe & Trevino, 1995, 1996). This led to the development of a two-factor theory of morality based on generalised traits and specific predictors (Burton, 1963). Others see cheating as deviant behaviour and explain it in terms of (a) deterrence theory, in which the probability and extent of punishment control behaviours; (b) rational choice theory, in which the probabilities of both rewards and punishments are included; (c) social bond theory, in which deviant behaviour is a result of the weakening of social bonds such as attachment, commitment, involvement and moral belief; and (d) social learning theory, in which deviant behaviour is reinforced in primary groups (see Michaels & Miethe, 1989).

A small number of studies involved observational or experimental findings. Observational studies usually involved (a) some form of surreptitious observation in which students have the opportunity to cheat; or (b) determining the overlap in errors of adjacent students in an exam with the overlap in errors of non-adjacent students; or (c) a randomised response technique which invites a binary response from a student (see Chaudhuri & Mukerjee, 1988 for a description of this method). Experimental methods involve manipulations, such as the examination of cheating on sex-appropriate tasks involving 11 very difficult and 4 easy questions (Lobel, 1993) or cheating under high- and low- risk conditions (Leming, 1980) or cheating by copying assignments across semesters (Karlins, Michaels, Freilinger & Walker, 1989). Observational and experimental manipulations are not included in this review.

Self-reports of cheating

Given the sensitivity of dishonest behaviours, most recent studies of cheating have relied upon survey methods involving anonymous self-report. Doubts about the credibility of this method have been noted (Bushway & Nash, 1977, p. 629; Spiller & Crown, 1995, p. 764). One, a feature of the self-reports is that they have yielded higher response rates from females in a number of studies (see McCabe & Trevino, 1997, p. 386). Secondly, useable response rates for some surveys have varied (e.g., 90% - Franklyn-Stokes and Newstead, 1995; 65% - Erickson & Smith, 1974) depending upon the circumstances under which the data was collected. Thirdly, the use of direct questioning methods may underestimate current class-specific cheating. For instance, Kerkvliett (1994) found that the proportion of students admitting to cheating using direct questioning (even when anonymous) was 0.259, compared with 0.419 using a randomized response technique. [In the randomized response technique, students generated a random number from their social security ID and this was categorised before they were required to answer truthfully. The probability of truthful responses can be determined based on comparisons of the assumed distribution of numbers and responses in the categories (see Chaudhuri & Mukerjee, 1988).

Despite its many limitations, the method of confidential and anonymous self-report has ethical and moral advantages. Hollinger and Lanza-Kaduce (1996) stated:

> … a self-administered survey provides the best opportunity to obtain detailed information from students about their academic dishonesty. It also avoids the ethical problems associated with contriving temptations to cheat and then deceiving students about it. Further, a survey instrument can be used to

> collect information efficiently about different forms of academic dishonesty across a variety of contexts. Surveys also permit students to remain anonymous. In general, confidential self-report surveys about minor forms of deviance among conventionally socialized individuals have been judged to be methodologically valid and reliable… (1996, p. 394)

This study reviews and evaluates the extent of any gender differences in academic cheating behaviours based on previous studies that used self-report data. The main research question was whether males reported higher rates of cheating than females, and a secondary question was whether any gender difference was consistent across assessment contexts. A meta-analysis was used to accumulate the results of previous studies because it offered a better representation of the relationship between gender and self-reports of cheating than can be provided by any one study. This meta-analysis used the effect size statistic, d (Cohen, 1988).

Literature search strategy

Computer-based searches of Psychological Abstracts and the Educational Resources Information Center (ERIC) databases were conducted only for published studies relating to gender and cheating. This was supplemented by checking the references cited for any further studies not located by the computer-based search. Twenty-one studies that reported on gender and cheating were located. These studies are summarised in Tables 1 and 2. Table 1 summarises those 14 studies that cited the proportion of men and women cheating and Table 2 lists the 21 studies used for the determination of an effect size.

TABLE 1: Proportion of male and female students who reported cheating

						Proportion cheating	
Authors	**Year**	**Sample**	**Females**	**Males**	**Country**	**Female**	**Male**
Ameen, Guffey & McMillan	1996	University	168	117	USA	0.518	0.624
Astin, Panos & Creager	1967	College	94,537	112,328	USA	0.165	0.241
Baldwin, Daugherty, Rowley & Schwarz	1996	Medical school	1,510	916	USA	0.302	0.468
Bowers	1964	University	2,568	2,810	USA	0.430	0.540
Burns, Davis, Hoshino & Miller	1998	University	77	151	Japan	0.449	0.370
		University	88	32	Sth Africa	0.193	0.438
		University	57	33	Sth Africa	0.404	0.545
Davis, Noble, Zak & Dreyer	1994	University	39	10	Aust	0.510	0.600
			1,478	675	USA	0.760	0.790
Erickson & Smith	1974	College students	68	50	USA	0.059	0.160
Huss, Curnyn, Roberts, Davis et al.	1993	College students	142	78	USA	0.730	0.770
Schab	1969	High school	580	835	USA	0.697	0.519
Schab	1980	High school	580	520	USA	0.952	0.899
Smith Ryan & Diggins	1972	College	68	44	USA	0.970	0.910
Thorpe, Pittenger & Reed	1999	University	81	57	USA	0.795	0.912
			124	48		0.621	0.729
Who's who amongst American high school students	1993	High school	1,429	528	USA	0.679	0.660
Who's who amongst American high school students	1994	High school	2,256	921	USA	0.722	0.632

Note: Includes largest proportion where multiple proportions are quoted

TABLE 2: Effect sizes and meta-analysis

Authors	Year	Level	Female	Male	School	College	Quiz	Unit test	Mid-term	Exam	Copy assgnmnt	Plagiarism	Other
Ameen, Guffey & McMillan	1996	University	168	117		0.211							
Antion & Michael	1983	Community college	84	64						0.262	-0.242		
Astin, Panos & Creager	1967	Community college	94,537	112,328						0.188			
Baird	1980	College	113	87			0	0.551	0.525	0.471	0.563		0.519
Baldwin, Daugherty, Rowley & Schwarz	1996	Medical school	916	1510	0.259								
					0.332								
						0.192							
						0.05							
Bowers	1964	University	2,568	2,810		0.221							
Burns, Davis, Hoshino & Miller	1998	University	77	151	0.045	0.077							
		University	88	32	0.247	0.266							
		University	57	33	0.137	0.041							
Davis & Ludvigson	1995	College	71	71		0.347							
			675	1,478	0								
Davis, Grover, Becker & McGregor	1992	College				0.281							
						0.337							
Davis, Noble, Zak & Dreyer	1994	College	39	10	0.147								
			1,478	675	0.067								
			39	10		0.452							
			1,478	675		0.205							
De Vries & Azjen	1971	College	73	73		0							
Erickson & Smith	1974	College	68	50						0.335			
Garfield, Cohen & Roth	1967	College	50	30		-0.324							
Huss, Curnyn, Roberts, Davis et al.	1993	College	142	78	0.088	0.194							
Schab	1969	High school	580	835							0.055	0.124	0.464

Schab	1980	High school	580	520			0.15	0.052	0.204
Smith Ryan & Diggins	1972	College	68	44		0.263			
Stern & Havlicek	1986	University	314	188		0.294			0.294
Thorpe, Pittenger & Reed	1999	University	81	57		0.466	0.472	0.342	0.321
			124	48		0.365	0.009	0.007	0.017
			81	57		0.377	0.429	0.238	0.076
			124	48		0.261	0.3	0.049	0.089
Who's who amongst American high school students	1993	High school	1,429	528	0.031		0.036	0.092	
Who's who amongst American high school students	1994	High school	2,256	921	0.042		0.178	0.091	

Note: Effect sizes in the high school and college columns are only for those studies that did not indicate the specific form of cheating; In Davis & Ludvigson (1995) the sample size refers to classes.

Criteria for including studies

Studies were included when they cited the proportion of males and females that cheated or they cited a statistic that could be converted to an effect size. Studies were excluded if they focused on the self-reported frequency of cheating, as this did not address the research question. As a result of this more stringent criterion, the reader should note significant variations in the studies included in this meta-analysis compared with that of Whitley, Nelson and Jones (1999).

A typical study using a single question was that of Smith, Ryan and Diggins (1972, p. 646). They asked "Have you ever cheated on an examination?" and 91% of men (N=44) and 97% of women (N=88) answered "Yes." In other studies, more than one question was asked and the findings from each question were used in the meta-analysis. For instance, Antion and Michael (1983) used two separate questions answered yes or no from the Marlowe-Crowne Social Desirability Scale: (a) "I have never cheated on a test" and (b) "I have never used somebody else's term paper." They reported correlations between sex and these questions of –0.09 and –0.04, respectively, for 148 community college students. These results were used independently as indicators of cheating on tests and cheating on term papers in the meta-analysis.

In some studies the proportion cheating was determined for various situations (e.g., tests, assignments, plagiarism). In those cases where more than one proportion was cited, the highest proportion cheating in any one context was used as an indicator of the extent of cheating in a high school or university. For example, the study by Who's Who Among American High School Students (1994) provided details of the proportions of students who copied someone else's homework (male = 63.2%; female = 72.2%), cheated on a quiz or test (male = 42.2%; female = 44.5%), or plagiarised part of an essay (male = 17.9%; female = 14.3%). The highest of the three proportions was used as an indicator of the extent of past cheating in high school. The separate questions were also used as indicators of cheating in particular contexts.

The study by Roth and McCabe (1995) used multiple questions and was typical of the reports not included in the meta-analysis. The dependent variable in their study was a composite measure. Students were asked how often they engaged in copying using crib notes, using unfair methods, or helping someone cheat. These behaviours were rated on a scale from 0 (never) to 5 (very often). The composite measure was the sum of the scale values. It focused on frequency of cheating and ratings and was included by Whitley, Nelson and Jones (1999) in their meta-analysis.

The study by Bonjean and McGee (1965) reported a larger percentage of males (0.716) than females (0.636) as actual or potential cheaters. It used a single measure but was not included. The report specified six cheating behaviours (seeking exam information from students, copying, collusion, lying about an absence, bringing information into an examination, purchasing a final exam). Students were asked "Have you or would you ever do this in the same situation?" Students answering "Yes" to any of the six situations were labeled as actual or potential violators. This study was excluded mainly because it contaminated past with potential future behaviour. Once again, Whitley, Nelson and Jones (1999) included this study in their meta-analysis.

Coding of studies

The studies that were selected were coded as follows (a) cohort (cheating in high school; or cheating in college and university); and (b) the specific context for cheating (e.g., cheating on tests or exams, plagiarism, copying assignments, lending work, and other forms of cheating). In this meta-analysis, gender differences in terms of the context were also examined.

The 21 studies included were published from 1964 (Bowers, 1964) through to 1999 (Thorpe, Pittenger & Reed, 1999). Two studies used community college students, 15 used college or university students, and four used high school students. The 14 studies listed in Table 1 cited the proportions of females and males who said they had cheated. In the meta-analysis (see Table 2) the studies varied in sample size from 49 to 206,865 and 56 out of the 64 effect sizes reported were from the USA. The total number of participants included for the meta-analysis were 108,358 females and 123,528 males.

Data analysis

The d statistic was used as the indicator of effect size. This is the standardised mean difference between men and women. Values of d are described as low (0.2), medium (0.5), or high (0.8) (Cohen, 1992). Differences in proportions were tested using the independent z-test and converted to chi-square with one-degree of freedom, then transformed to a correlation coefficient and an effect size. Studies such as Davis et al. (1992) reported only chi-square values for sub-samples and these were used to calculate a correlation and effect size. Studies that reported a probability such as $p<0.01$ were converted to two-tail z-scores and from there to chi-square, correlations and then effect sizes. In those instances where a study did not report a statistically significant gender difference, the effect size was categorised as zero following Whitley, Nelson and Jones (1999). There were varying numbers of participants in the studies analysed and the effect size was weighted by the size of each study. Where possible the results were checked against published findings; for example, the proportions in the Erickson and Smith (1974) article were used to recalculate the z-score quoted in the published paper (p. 109). A spreadsheet setting out the computations is available from the author upon request. The formulae for conversion of proportions into effect sizes are listed in Appendix A. In the following section, the results are discussed firstly in terms of the proportion of students who cheated and secondly in terms of the effect sizes for differences between males and females.

RESULTS

Proportion of male and female students cheating

The findings from earlier research confirmed that a large number of students had cheated. The overall proportions of female students cheating varied from a low of 0.05 to a high of 0.97 (median = 0.56) and for men, the proportion varied from 0.16 to 0.91 (median = 0.61). There was no statistical significant difference in the average proportions reported for males and females (t(34) = -0.58, ns). Accumulating the findings across the studies that reported both proportions and the actual number of males and females involved (N=226,003), showed that 21% of females and 26% of males had cheated. If the extremely large sample in the study by Astin et al. (1967) is excluded, then the proportions increase dramatically to 60% for both males and females. These studies showed a wide dispersion of findings (see Figure 1 for the distribution of proportions).

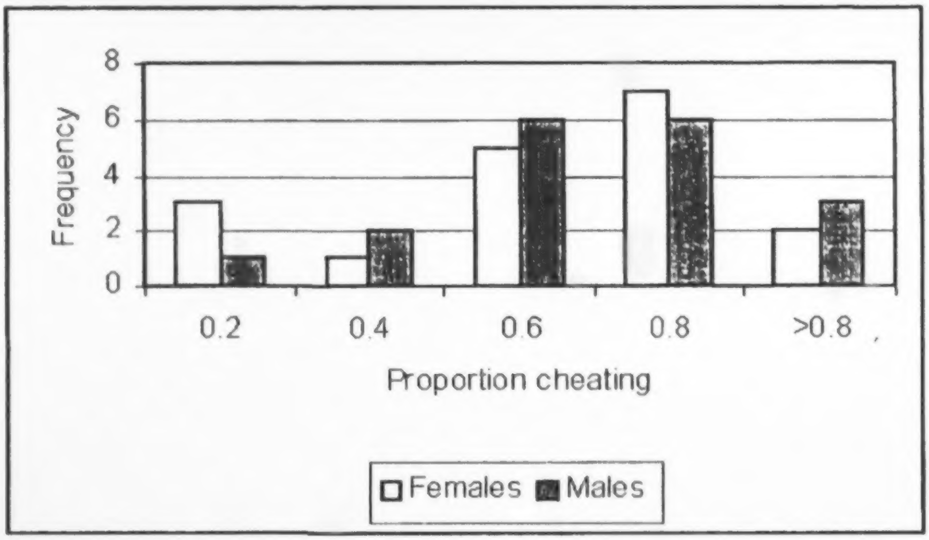


FIGURE 1: Distribution of the proportions of males and females cheating (N=18 studies).

Effect sizes

The mean effect size between males and females for cheating in high schools was obtained from nine studies and was 0.14 (SE=0.03), and the mean effect size from 15 studies that considered cheating in college or university was slightly higher at 0.0.17 (SE=0.04). The overall effect size obtained from all studies (but using only the largest effect size where more than one was quoted) was 0.23 (SE=0.02). The effect sizes were not uniform but for the most part could be characterised as around 0.2 (see Table 3 for a listing of the average effect sizes for various assessment contexts).

TABLE 3: Mean effect sizes

Assessment contexts	N	Mean effect (includes multiple observations)	H	chi-square
High school	9	0.172	38.7	ns
College	13	0.181	34.2	ns
All exams, quizzes, unit and mid-term tests	15	0.186	47.5	ns
All exams etc (excl. Astin et al.)	14	0.141	44.4	ns
Copy assignment	10	0.132	29	ns

Plagiarism	8	0.128	6.3	df=7, p<.05
Lent work, other	8	0.315	19.5	ns
Overall effect size	29	0.191	96.1	ns

Note: Overall effect size used only the largest effect size from each group where more than one effect size was available.

The reader should note that the effect sizes for exams, quizzes, mid-term or unit tests were combined. This means that some studies provided more than one effect size and that the average effect size was not based on the ideal of independent observations. The average effect size for tests and exams was 0.186 and 0.141 if the results from the very large study by Astin et al. (1967) were excluded. This compared with 0.132 for the copying of assignments. The results from averaging the effect sizes from studies involving plagiarism involved eight effect sizes and had an average of 0.128. The remaining assessment context (lending work or other forms of cheating) involved eight effect sizes with an overall effect size of 0.315. Indeed, this was the largest effect size reported and points to an area that is worthy of further analysis. All of these effect sizes were plagued by problems of heterogeneity, and it is likely that a consistent average effect size has not been determined. The only exception appears to be that of the eight effect sizes cited for plagiarism, which appear to be relatively homogeneous and the net gender effect is very low.

The largest effect size of 0.332 for high school students came from a retrospective self-report cited by Baldwin et al. (1996) in their survey of medical school students. In fact, not one of the effect sizes for high school students came from a survey of that cohort. In contrast, the highest effect size for college students was 0.452 from a study of only 49 college students by Davis et al. (1994).

On exams, the largest effect size of 0.471 was reported by Baird (1980) from 200 college students; this study also reported the largest effect size (0.563) for copying. In fact, this was the largest of the 65 effect sizes calculated in Table 2. There were only two negative effect sizes; –0.242 reported for copying assignments by Antion and Michael (1983) and –0.324 reported for 50 female and 30 male college students by Garfield et al. (1967).

DISCUSSION

The findings from this review and evaluation of earlier studies indicated that substantial proportions of males and females engaged in cheating at high school and that substantial proportions continued cheating in college or university. It is a major concern for those involved in assessment that so many admitted cheating in some form or another. If we exclude the study of Astin et al., then the proportion of males or females admitting to cheating was 60% for both groups, reducing to 21% and 26% respectively for males and females if this study was included. Even these lower percentages (21% and 26%) must still be cause for some action.

The astute reader might note that in 12 of the 18 studies in Table 1 the proportion of males cheating outnumbered the proportion of females cheating. This is a case of Simpson's Paradox, where inferences from large data sets are often the opposite of inferences from smaller sets. The paradox is caused by data from unequal sized groups being accumulated inappropriately into one large group and may also indicate the effect of some other intervening factor (e.g., sampling influences, methodology, disproportionate responding between males and females).

There was some support, however, for the view that there were small differences in the proportion of males and females cheating. The overall effect sizes for high school and college cohorts were low ($d \approx 0.2$). The significant heterogeneity in the effect sizes means that our estimates are still sporadic, diverse, and have not yet yielded consistent findings.

These low effect sizes for male-female differences were consistent with the small effect size reported by Whitley, Nelson and Jones (1999). We used seven reports containing eight proportions that were not cited by Whitley, Nelson and Jones (1999) as well as ten studies with 30 effect sizes not contained in their article. Moreover, the basis of the classification of studies included in this paper varied from the approach outlined by Whitley, Nelson and Jones (1999).

Some limitations of this study arise from the fact that for the most part it focused on only two of the three forms of cheating identified by Cizek (1999), namely cheating by taking, giving, or receiving information from others and possibly cheating through the use of forbidden materials or information. It is not clear to what extent cheating by circumventing the process of assessment was covered. In addition, this paper has not covered observational and experimental approaches to cheating, and findings from these studies may validate or qualify the findings obtained.

A second limitation of this review is that cheating in contexts such as primary schooling, and adult and vocational education was not covered. Moreover, the effect sizes in the meta-analysis relied largely on the retrospective reports from college students on the extent of their high school cheating. These may not be accurate and may also be influenced by perceptions of what it means to cheat. In addition, almost all of the self-report studies were based on samples in the United States. Even a cursory knowledge of cheating practices worldwide indicates that the full dimensions of cheating across cultures might not be evident in the self-reports - mainly from the United States - that were reviewed in this paper.

The use of the confidential, anonymous, and private self-report surveys does have some advantages but it was also clear that females outnumbered male respondents in many of the studies. This occurred for 14 out of the 18 studies of proportions and for 20 out of the 25 groups with effect sizes. It may be that the propensity of females to answer such surveys on cheating is in some way linked with the extent of cheating among males or females. For instance a lower response rate may mask the prevalence of male cheating. The disproportionate participation of males and females is especially evident in the Who's Who (1993, 1994) studies of the attitudes of leaders in high schools. These studies had low response rates to a mail survey (3,177 out of 8,000 in 1994), and the final group comprised only 29% males. This limits seriously any conclusions that may be drawn about the extent of cheating. Finally, the findings from this review indicate the prevalence of cheating in a group, whereas an indication of the specific incidence may be more helpful for educators. For instance, Kerkvliett and Sigmund (1994) reported that only 1.9% of students admitted to cheating in a particular class.

This study confirms that cheating is a major educational problem and one that is likely to devalue assessment findings at all levels. Small differences between males and females were evident, but the effect of these differences was quite low. Both male and female students have cheated in large numbers, and unfortunately this affects many aspects of teaching, learning, and assessment and can disadvantage honest students.

Acknowledgments

An earlier version of this paper was presented at the International Conference on Measurement and Evaluation in Education, Universiti Sains Malaysia, Penang, November 2001. The helpful comments of two anonymous reviewers are gratefully acknowledged.

References

*Ameen, E.C., Guffey, D.M. & McMillan, J.J. (1996). Gender differences in determining the ethical sensitivity of future accounting professionals. *Journal of Business Ethics, 15*, 591-597.

*Antion, D.L. & Michael, W.B. (1983). Short-term predictive validity of demographic, affective, personal and cognitive variables in relation to two criterion measures of cheating behaviors. *Educational and Psychological Measurement, 43*, 467-482.

*Astin, A.W., Panos, R.J. & Creager, J.A. (1967). National norms for entering college freshmen – Fall 1966. *American Council on Education Research Reports*, 2(1).

*Baird, J.S. (1980). Current trends in college cheating. *Psychology in the Schools, 17*, 515-522.

*Baldwin, D.C., Daugherty, S.R., Rowley, B.D. & Schwarz, M.R. (1996). Cheating in medical school: A survey of second-year students at 31 schools. *Academic Medicine, 71*, 267-273.

Black, D.B. (1962). The falsification of reported examination marks in a senior university education course. *Journal of Educational Sociology, 35*, 346-54.

Bonjean, C.M. & McGee, R. (1965). Scholastic dishonesty among undergraduates in differing systems of social control. *Sociology of Education, 28*, 127-137.

*Bowers, W.J. (1964). *Student dishonesty and its control in college*. New York: Bureau of Applied Social Research, Columbia University.

*Burns, S.R., Davis, S.F., Hoshino, J. & Miller, R.L. (1998). Academic dishonesty: A delineation of cross-cultural patterns. *College Student Journal, 32*, 590-596.

Burton, R.V. (1963). Generality of honesty reconsidered. *Psychological Review, 70*, 481-499.

Bushway, A. & Nash, W.R. (1977). School cheating behavior. *Review of Educational Research, 47*, 623-632.

Chaudhuri, A., & Mukerjee, R. (1988). *Randomized response: Theory and techniques.* New York: Marcel Dekker.

Cizek, G.J. (1999). *Cheating on tests: How to do it, detect it and prevent it*. Mahwah, NJ: Lawrence Erlbaum.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd Ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

*Davis, S.F., Grover, C.A., Becker, A.H. & McGregor, L.N. (1992). Academic dishonesty: Prevalence, determinants, techniques and punishments. *Teaching of Psychology, 19*, 16-20.

*Davis, S.F. & Ludvigson, H.W. (1995). Additional data on academic dishonesty and a proposal for remediation. *Teaching of Psychology, 22*, 119-122.

*Davis, S.F., Noble, L.M., Zak, E.N. & Dreyer, K.K. (1994). A comparison of cheating and learning/grade orientation in American and Australian college students. *College Student Journal, 28*, 353-6.

*De Vries, D.L. & Azjen, I. (1971). The relationship of attitudes and normative beliefs to cheating in college. *The Journal of Social Psychology, 83*, 199-207.

*Erickson, M.L. & Smith, W.B. (1974). On the relationship between self-reported and actual deviance: An empirical test. *Humboldt Journal of Social Relations, 10,* 106-113.

Franklyn-Stokes, A. & Newstead, S.E. (1995). Undergraduate cheating: Who does what and why? *Studies in Higher*

Education, 20, 159-172.

Gabertson, K.B. (1997). Academic dishonesty among nursing students. *Nursing Forum, 32*(3), 14.

*Garfield, S.J., Cohen, H.A. & Roth, R.M. (1967). A correlative study of cheating in college students. *The Journal of Educational Research, 61*, 172-173.

Good, T.L., Nichols, S.L. & Sabers, D.L. (1999). Underestimating youth's commitment to schools and society: Toward a more differentiated view. *Social Psychology of Education, 3*, 1-39.

Graham, M.A., Monday, J., O'Brien, K. & Steffen, S. (1994). Cheating at small colleges: An examination of student and faculty attitudes and behaviors. *Journal of College Student Development, 35*, 255-260.

Haines, V.J., Dickhoff, G.M., LaBeff, E.E. & Clark, R.E. (1986). College cheating: Immaturity, lack of commitment and the neutralizing attitude. *Research in Higher Education, 25*, 342-354.

Hartshorne, H. & May, M.A. (1928). *Studies in deceit.* New York: McMillan.

Hollinger, R.C. & Lanza-Kaduce, L. (1996). Academic dishonesty and the perceived effectiveness of countermeasures: An empirical survey of cheating at a major public university. *NASPA Journal, 33*, 292-306.

Horne, F.W. (1965). Attitudes on cheating of high school students. *Journal of the National Association of Women Deans and Counselors, 28*, 102-106.

*Huss, M.T., Curnyn, J.P., Roberts, S.L., Davis, S.F., Yandell, L. & Giordano, P. (1993). Hard driven but not dishonest: Cheating and the Type A personality. *Bulletin of the Psychonomic Society, 31*, 429-430.

Kerkvliett, J. (1994). Cheating by economics students: A comparison of survey results. *Journal of Economic Education, 25*, 121-133.

Kerkvliett, J. & Sigmund, C.L. (1999). Can we control cheating in the classroom? *The Journal of Economic Education, 30*, 331.

Karlins, M., Michaels, C., Freilinger, P. & Walker, H. (1989). Sex difference sin academic dishonesty: College cheating in a management course. *Journal of Education for Business, 65*, 31-33.

Leming, J.S. (1980). Cheating behaviour, subject variables and components of the internal-external scale under high and low risk conditions. *Journal of Educational Research, 74*, 83-87.

Lobel, T.E. (1993). Gender diffrences in adolescents' cheating behavior: An interactional model. *Personality and Individual Differences*, 14, 275-277.

McCabe, D.L. & Trevino, L.K. (1995). Cheating among business students: A challenge for business leaders and educators. *Journal of Management Education, 19*, 205-218.

McCabe, D.L. & Trevino, L.K. (1996). What we know about cheating in college. *Change*, January/February, 29-33.

McCabe, D.L. & Trevino, L.K. (1997). Individual and contextual influences on academic dishonesty: A multicampus investigation. *Research* *in Higher Education*, 38, 379-396.

Michaels, J.W. & Miethe, T.D. (1989). Applying theories of deviance to academic cheating. *Social Science Quarterly, 70*, 870-885.

Newstead, S.E., Franklyn-Stokes, A. & Armstead, P. (1996). Individual differences in student cheating. *Journal of Educational Psychology, 88*, 229-241.

Roth, N.L. & McCabe, D.L. (1995). Communication strategies for addressing academic dishonesty. *Journal of College Student Development, 36*, 531-541.

*Schab, F. (1969). Cheating in high school: Differences between the sexes. *National Association of Women Deans and Counselors,33*, 39-42.

*Schab, F. (1980). Cheating in high school: Differences between the sexes (Revisited). *Adolescence, 15*, 959-965.

Sierles, F., Hendrickx, I., & Circle, S. (1980). Cheating in medical school. *Journal of Medical Education, 55*, 124-125.

*Smith, C.P., Ryan, E.R. & Diggins, D.R. (1972). Moral decision making: Cheating on examinations. *Journal of Personality, 40*, 640-660.

Spiller, S. & Crown, D.F. (1995). Changes over time in academic dishonesty at the collegiate level. *Psychological Reports, 76*, 763-768.

*Stern, E.B. & Havlicek, L. (1986). Academic misconduct: Results of faculty and undergraduate student surveys. *Journal of Allied Health*, May 1986, 129-142.

*Thorpe, M.T., Pittenger, D.J. & Reed, B.D. (1999). Cheating the researcher: A study of the relation between personality measures and self-reported cheating. *College Student Journal, 33*, 49.

Whitley, B.E., Jr. (1998). Factors associated with cheating among college students: A review. *Research in Higher Education, 39*, 235-274.

Whitley, B.E., Jr., Nelson, A.B. & Jones, C.J. (1999). Gender differences in cheating attitudes and classroom cheating behavior: A meta-analysis. *Sex Roles: A Journal of Research*, 657 .

*Who's Who among American High School Students (1993). *Attitudes and Opinions from the Nation's High Achieving Teens. 24th Annual Survey of High Achievers*. Lake Forest, IL.: Author.

*Who's Who among American High School Students (1994). *Attitudes and Opinions from the Nation's High Achieving Teens. 25th Annual Survey of High Achievers*. Lake Forest, IL.: Author.

* references with an asterisk are used in the meta-analysis of results

APPENDIX A

Formulae for conversion of proportions to effect sizes

Z-score $z = (p_1 - p_2)/ \text{SQRT} ((p*q)*((1/n_1) + (1/n_2)))$

Chi-square $c^2 = z^2$

Correlation $r = \text{SQRT} (c/ (n_1 + n_2))$

Effect size $d = 2r/(\text{SQRT}(1 - r^2))$

Descriptors: Meta analysis; Sex differences; Error, Cheating; Student Behavior; Student Evaluation; Academic Misconduct

```
ADODB.Recordset error '800a0e78'

Operation is not allowed when the object is closed.

/pare/getvn.asp, line 179
```

Home | Articles | Subscribe | Review | Policies

Volume: 8 7 6 5 4 3 2 1

Practical Assessment,
Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ☐ Go

Copyright 2002, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Osborne, Jason (2002). Notes on the use of data transformations. *Practical Assessment, Research & Evaluation*, 8(6). Available online: http://ericae.net/pare/getvn.asp?v=8&n=6.
This paper has been viewed 3105 times since 5/30/02.

▸ Find similar papers in
ERICAE Full Text Library
Pract Assess, Res & Eval
ERIC RIE & CIJE 1990-
ERIC On-Demand Docs

▸ Find articles in ERIC written by
Osborne, Jason

Notes on the use of data transformations.

Jason W. Osborne, Ph.D
North Carolina State University

Data transformations are commonly used tools that can serve many functions in quantitative analysis of data. The goal of this paper is to focus on the use of three data transformations most commonly discussed in statistics texts (square root, log, and inverse) for improving the normality of variables. While these are important options for analysts, they do fundamentally transform the nature of the variable, making the interpretation of the results somewhat more complex. Further, few (if any) statistical texts discuss the tremendous influence a distribution's minimum value has on the efficacy of a transformation. The goal of this paper is to promote thoughtful and informed use of data transformations.

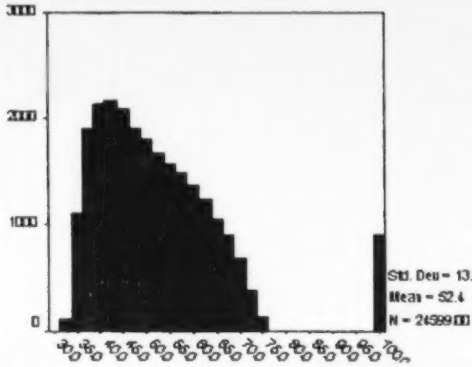
Data transformations are the application of a mathematical modification to the values of a variable. There are a great variety of possible data transformations, from adding constants to multiplying, squaring or raising to a power, converting to logarithmic scales, inverting and reflecting, taking the square root of the values, and even applying trigonometric transformations such as sine wave transformations. The goal of this paper is to begin a discussion of some of the issues involved in data transformation as an aid to researchers who do not have extensive mathematical backgrounds, or who have not had extensive exposure to this issue before, particularly focusing on the use of data transformation for normalization of variables.

Data transformation and normality

Many statistical procedures assume that the variables are normally distributed. A significant violation of the assumption of normality can seriously increase the chances of the researcher committing either a Type I or II error (depending on the nature of the analysis and the non-normality). However, Micceri (1989) points out that true normality is exceedingly rare in education and psychology. Thus, one reason (although not the only reason)

BEST COPY AVAILABLE

researchers utilize data transformations is improving the normality of variables. Additionally, authors such as Zimmerman (e.g., 1995, 1998) have pointed out that non-parametric tests (where no explicit assumption of normality is made) can suffer as much, or more, than parametric tests when normality assumptions are violated, confirming the importance of normality in all statistical analyses, not just parametric analyses.



There are multiple options for dealing with non-normal data. First, the researcher must make certain that the non-normality is due to a valid reason (real observed data points). Invalid reasons for non-normality include things such as mistakes in data entry, and missing data values not declared missing. Researchers using NCES databases such as the National Education Longitudinal Survey of 1988 will often find extreme values that are intended to be missing. In Figure 1 we see that the Composite Achievement Test scores variable (BY2XCOMP) ranges from about 30 to about 75, but also has a group of missing values assigned a value of 99. If the researcher fails to remove these the skew for this variable is 1.46, but with the missing values appropriately removed, skew drops to 0.35, and thus no further action is needed. These are simple to remedy through correction of the value or declaration of missing values.

However, not all non-normality is due to data entry error or non-declared missing values. Two other reasons for non-normality are the presence of outliers (scores that are extreme relative to the rest of the sample) and the nature of the variable itself. There is great debate in the literature about whether outliers should be removed or not. I am sympathetic to Judd and McClelland's (1989) argument that outlier removal is desirable, honest, and important. However, not all researchers feel that way (c.f. Orr, Sackett, and DuBois, 1991). Should a researcher remove outliers and find substantial non-normality, or choose not to remove outliers, data transformation is a viable option for improving normality of a variable. It is beyond the scope of this paper to fully discuss all options for data transformation. This paper will focus on three of the most common data transformations utilized for improving normality discussed in texts and the literature: square root, logarithmic, and inverse transformations. Readers looking for more information on data transformations might refer to Hartwig and Dearing (1979) or Micceri (1989).

How does one tell when a variable is violating the assumption of normality?

There are several ways to tell whether a variable is substantially non-normal. While researchers tend to report favoring "eyeballing the data," or visual inspection (Orr, Sackett, and DuBois, 1991), researchers and reviewers often are more comfortable with a more objective assessment of normality, which can range from simple examination of skew and kurtosis to examination of P-P plots (available through most statistical software packages) and inferential tests of normality, such as the Kolmorogov-Smirnov test (and adaptations of this test—researchers wanting more information on the K-S test and other similar tests should consult the manual for their software as well as Goodman (1954), Lilliefors (1967), Rosenthal (1968), and Wilcox (1997), probably in that order). These can be useful to a researcher needing to know whether a variable's distribution is significantly different from a normal (or other) distribution.

Notes on the mathematics of these data transformations

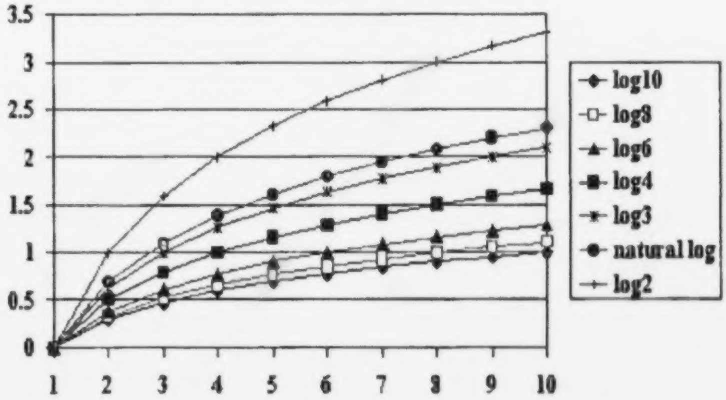
While many researchers in the social sciences are well-trained in statistical methods, not many of us have had significant mathematical training, or if we have, it has often been long forgotten. This section is intended to give a brief refresher on what really happens when one applies a data transformation.

Square root transformation. Most readers will be familiar with this procedure-- when one applies a square root transformation, the square root of every value is taken. However, as one cannot take the square root of a negative number, if there are negative values for a variable a constant must be added to move the minimum value of the distribution above 0, preferably to 1.00 (the rationale for this assertion is explained below). Another important point is that numbers of 1.00 and above behave differently than numbers between 0.00 and 0.99. The square root of numbers above 1.00 always become smaller, 1.00 and 0.00 remain constant, and numbers between 0.00 and 1.00 become larger (the square root of 4 is 2, but the square root of 0.40 is 0.63). Thus, if you apply a square root to a continuous variable that contains values between 0 and 1 as well as above 1, you are treating some numbers differently than others, which is probably not desirable in most cases.

Log transformation(s). Logarithmic transformations are actually a class of transformations, rather than a single transformation. In brief, a logarithm is the power (exponent) a base number must be raised to in order to get the original number. Any given number can be expressed as y to the x power in an infinite number of ways. For example, if we were talking about base 10, 1 is 10^0, 100 is 10^2, 16 is $10^{1.2}$, and so on. Thus, $\log_{10}(100)=2$ and $\log_{10}(16)=1.2$. However, base 10 is not the only option for log transformations. Another common option is the Natural Logarithm, where the constant *e* (2.7182818) is the base. In this case the natural log 100 is 4.605. As the logarithm of any negative number or number less than 1 is undefined, if a variable contains values less than 1.0 a constant must be added to move the minimum value of the distribution, preferably to 1.00.

There are good reasons to consider a range of bases (Cleveland (1984) argues that base 10, 2, and *e* should always be considered at a minimum). For example, in cases where there are extremes of range base 10 is desirable, but when there are ranges that are less extreme, using base 10 will result in a loss of resolution, and using a lower base (*e* or 2) will serve (higher bases tend to pull extreme values in more drastically than lower bases). Figure 1 graphically presents the different effects of using different log bases. Readers are encouraged to consult Cleveland (1984).

Figure 1.
The Effect of log base on the efficacy of transformations.



Inverse transformation. To take the inverse of a number (*x*) is to compute 1/*x*. What this does is essentially make very small numbers very large, and very large numbers very small. This transformation has the effect of reversing the order of your scores. Thus, one must be careful to reflect, or reverse the distribution prior to applying an inverse transformation. To reflect, one multiplies a variable by -1, and then adds a constant to the distribution to bring the minimum value back above 1.0. Then, once the inverse transformation is complete, the ordering of the values will be identical to the original data.

In general, these three transformations have been presented in the relative order of power (from weakest to most powerful). However, it is my preference to use the minimum amount of transformation necessary to improve normality.

Positive vs. Negative Skew. There are, of course, two types of skew: positive and negative. All of the above-mentioned transformations work by compressing the right side of the distribution more than the left side. Thus, they are effective on positively skewed distributions. Should a researcher have a negatively skewed distribution, the researcher must reflect the distribution, add a constant to bring it to 1.0, apply the transformation, and then reflect again to restore the original order of the variable.

Issues surrounding the use of data transformations

Data transformations are valuable tools, with many benefits. However, they should be used appropriately, in an informed manner. Too many statistical texts gloss over this issue, leaving researchers ill-prepared to utilize these tools appropriately. All of the transformations examined here reduce non-normality by reducing the relative spacing of scores on the right side of the distribution more than the scores on the left side.

However, the very act of altering the relative distances between data points, which is how these transformations improve normality, raises issues in the interpretation of the data. If done correctly, all data points remain in the same relative order as prior to transformation. This allows researchers to continue to interpret results in terms of increasing scores. However, this might be undesirable if the original variables were meant to be substantively interpretable (e.g., annual income, years of age, grade, GPA), as the variables become more complex to interpret due to the curvilinear nature of the transformations. Researchers must therefore be careful when interpreting results based on transformed data. This issue is illustrated in Figure 2 and Table 1.

Figure 2.
The Effect of Transformations on Variables.

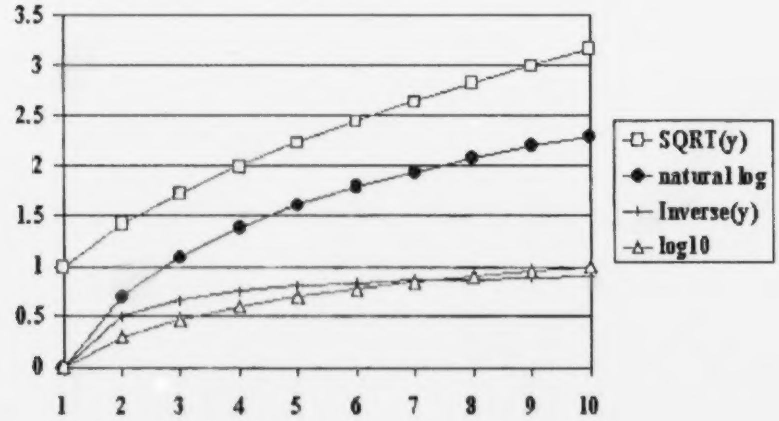


Table 1: Effects of various transformations on variables

Original Y	**1.00**	**2.00**	**3.00**	**4.00**	**5.00**	**6.00**	**7.00**	**8.00**	**9.00**	**10.00**
SquareRoot(Y)	1.00	1.41	1.73	2.00	2.24	2.45	2.65	2.83	3.00	3.16
gap		0.41	0.32	0.27	0.24	0.21	0.20	0.18	0.17	0.16
% reduction	0.00	29.29	42.26	50.00	55.28	59.18	62.20	64.64	66.67	68.38
Log10 (Y)	0.00	0.30	0.48	0.60	0.70	0.78	0.85	0.90	0.95	1.00
gap		0.30	0.18	0.12	0.10	0.08	0.07	0.06	0.05	0.05
% reduction	100.00	84.95	84.10	84.95	86.02	87.03	87.93	88.71	89.40	90.00
Reflected Inverse(Y)	0.00	0.50	0.67	0.75	0.80	0.83	0.86	0.88	0.89	0.90
gap		0.50	0.17	0.08	0.05	0.03	0.02	0.02	0.01	0.01
% reduction	100.00	75.00	77.78	81.25	84.00	86.11	87.76	89.06	90.12	91.00
Original Y	**11.00**	**12.00**	**13.00**	**14.00**	**15.00**	**16.00**	**17.00**	**18.00**	**19.00**	**20.00**
SquareRoot(Y)	3.32	3.46	3.61	3.74	3.87	4.00	4.12	4.24	4.36	4.47
gap		0.15	0.14	0.14	0.13	0.13	0.12	0.12	0.12	0.11
% reduction	69.85	71.13	72.26	73.27	74.18	75.00	75.75	76.43	77.06	77.64
Log10 (Y)	1.04	1.08	1.11	1.15	1.18	1.20	1.23	1.26	1.28	1.30
gap		0.04	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.02
% reduction	90.53	91.01	91.43	91.81	92.16	92.47	92.76	93.03	93.27	93.49
Reflected Inverse(Y)	0.91	0.92	0.92	0.93	0.93	0.94	0.94	0.94	0.95	0.95
gap		0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
% reduction	91.74	92.36	92.90	93.37	93.78	94.14	94.46	94.75	95.01	95.25
Original Y	**100.00**	**101.00**	**102.00**	**103.00**	**104.00**	**105.00**	**106.00**	**107.00**	**108.00**	**109.00**
SquareRoot(Y)	10.00	10.05	10.10	10.15	10.20	10.25	10.30	10.34	10.39	10.44
gap		0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
% reduction	90.00	90.05	90.10	90.15	90.19	90.24	90.29	90.33	90.38	90.42
Log10 (Y)	2.00	2.00	2.01	2.01	2.02	2.02	2.03	2.03	2.03	2.04
gap		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
% reduction	98.00	98.02	98.03	98.05	98.06	98.08	98.09	98.10	98.12	98.13
Reflected Inverse(Y)	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
gap		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
% reduction	99.01	99.02	99.03	99.04	99.05	99.06	99.07	99.07	99.08	99.09

While the original variable has equal spacing between values in Figure 2 (the X axis represents the original values), the other three lines depict the curvilinear nature of the transformations. The quality of the transformed variable is different from the original variable. If a variable with those qualities were subjected to a square root transformation, where the variable's old values were {0, 1, 2, 3, 4} the new values are now {0, 1, 1.41, 1.73, 2}—the intervals are no longer equal between successive values. The examples presented in Table 1 elaborate on this point. It quickly becomes evident that these transformations change the relative distance between adjacent values that were previously equidistant (assuming interval or ratio measurement). In the non-transformed variable, the distance between values would be an equal 1.0 distance between each increment (1, 2, 3, etc.). However, the action of the transformations dramatically alters this equal spacing. For example, where the original distance between 1 and 2 had been 1.0, now it is 0.41, 0.30, or 0.50, depending on the transformation. Further, while the original distance between 19 and 20 had been 1.0 in the original data, it is now 0.11, 0.02, or 0.00, depending on the transformation. Thus, while the order of the variable has been retained, order is all that has been maintained. The equal spacing of the original variable has been eliminated. If a variable had been measured on interval or ratio scales, it has now been reduced to ordinal (rank) data. While this might not be an issue in some cases, there are some statistical procedures that assume interval or ratio measurement scales.

Does the minimum value of a distribution influence the efficacy of a transformation?

For researchers with a strong mathematical or statistical background, the points made in this section are self-evident. However, over the years many of my students and colleagues have helped me to realize that to many researchers this point is not self-evident; further, it is not explicitly discussed in many statistical texts.

First, note that adding a constant to a variable changes only the mean, not the standard deviation or variance, skew, or kurtosis. However, the size of the constant and the place on the number line that the constant moves the distribution to can influence the effect of any subsequent data transformations. As alluded to above, it is my opinion that researchers seeking to utilize any of the above-mentioned data transformations should first move the distribution so its leftmost point (minimum value) is anchored at 1.0.

This is due to the differential effects of the transformations across the number line. All three transformations will have the greatest effect if the distribution is anchored at 1.0, and as the minimum value of the distribution moves away from 1.0 the effectiveness of the transformation diminishes dramatically.

Recalling that these transformations improve normality by compressing one part of a distribution more than another, the data presented in Table 1 illustrates this point. For all three transformations, the gap between 1 and 2 is much larger than between 9 and 10 (0.41, 0.30, and 0.50 vs. 0.16, 0.05, 0.01). Across this range, the transformations are having an effect by compressing the higher numbers much more than the lower numbers. This does not hold once one moves off of 1.0, however. If one had a distribution achored at 10 and ranging to 20, the gap between 10 and 11 (0.15, 0.04, 0.01) is not that much different than the gaps between 19 and 20 (0.11, 0.02, 0.00). In a more extreme example, the difference between 100 and 101 is almost the same as between 108 and 109.

In order to demonstrate the effects of minimum values on the efficacy of transformations, data were drawn from the National Education Longitudinal Survey of 1988. The variable used represented the number of undesirable things (offered drugs, had something stolen, threatened with violence, etc.) that had happened to a student, which was created by the author for another project. This variable ranged from 0 to 6, and was highly skewed, with 40.4% reporting none of the events occurring, 34.9% reporting only one event, and less than 10% reporting more than two of the events occurring. The initial skew was 1.58, a substantial deviation from normality, making this variable a good candidate for transformation. The relative effects of transformations on the skew of this variable are presented in Table 2.

Table 2: Variable skew as a function of the minimum score of a distribution

	Original Variable	Min = 1	Min = 2	Min = 3	Min = 5	Min = 10	Min = 100
Square Root	1.58	0.93	1.11	1.21	1.31	1.42	1.56
Log(10)	1.58	0.44	0.72	0.88	1.07	1.27	1.54
Inverse	1.58	0.12	0.18	0.39	0.67	1.00	1.50

As the results indicate, all three types of transformations worked very well on the original distribution, anchored at a minimum of 1. However, the efficacy of the transformation quickly diminished as constants were added to the distribution. Even a move to a minimum of 2 dramatically diminished the effectiveness of the transformation. Once the minimum reached 10, the skew was over 1.0 for all three transformations, and at a minimum of 100 the skewness was approaching the original, non-transformed skew in all three cases. These results highlight the importance of the minimum value of a distribution should a researcher intend to employ data transformations on that variable.

These results should also be considered when a variable has a range of, say 200-800, as with SAT or GRE scores where non-normality might be an issue. In cases where variables do not naturally have 0 as their minimum, it might be useful to subtract a constant to move the distribution to a 0 or 1 minimum.

Conclusions and other directions

Unfortunately, many statistical texts provide minimal instruction on the utilization of simple data transformations for the purpose of improving the normality of variables, and coverage of the use of other transformations or for uses other than improving normality is almost non-existent. While seasoned statisticians or mathematicians might intuitively understand what is discussed in this paper, many social scientists might not be aware of some of these issues.

The first recommendation from this paper is that researchers always examine and understand their data *prior to* performing those long-awaited analyses. To do less is to slight your data, and potentially draw incorrect conclusions.

The second recommendation is to know the requirements of the data analysis technique to be used. As Zimmerman (e.g., 1995, 1998) and others have pointed out, even non-parametric analyses, which are generally thought to be "assumption-free" can benefit from examination of the data.

The third recommendation is to utilize data transformations with care—and never unless there is a clear reason. Data transformations can alter the fundamental nature of the data, such as changing the measurement scale from interval or ratio to ordinal, and creating curvilinear relationships, complicating interpretation. As discussed above, there are many valid reasons for utilizing data transformations, including improvement of normality, variance stabilization, conversion of scales to interval measurement (for more on this, see the introductory chapters of Bond and Fox (2001), particularly pages 17-19).

The fourth recommendation is that, if transformations are to be utilized, researchers should ensure that they anchor the variable at a place where the transformation will have the optimal effect (in the case of these three, I argue that anchor point should be 1.0).

Beyond that, there are many other issues that researchers need to familiarize themselves with. In particular, there are several peculiar types of variables that benefit from attention. For example, proportion and percentage variables (e.g., percent of students in a school passing end-of-grade tests) and count variables of the type I presented above (number of events happening) tend to violate several assumptions of analyses and produce highly-skewed distributions. While beyond the scope of this paper, these types of variables are becoming increasingly common in education and the social sciences, and need to be dealt with appropriately. The reader interested in these issues should refer to sources such as Bartlett (1947) or Zubin (1935), or other, more modern sources that deal with these issues, such as Hopkins (2002, available at http://www.sportsci.org/resource/stats/index.html). In brief, when using count variables researchers should use the square root of the counts in the analyses, which takes care of count data issues in most cases. Proportions require an arcsine-root transformation. In order to apply this transformation, values must be between 0 and 1. A square root of the values is taken, and the inverse sine (arcsine) of that number is the resulting value. However, in order to use this variable in an analysis, each observation must be weighted by the number in the denominator of the proportion.

References

Baker, G. A. (1934). Transformation of non-normal frequency distributions into normal distributions. *Annals of Mathematical Statistics, 5,* 113-123.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences.* Mahwah, NJ: Lawrence Erlbaum.

Bartlett, M. S., (1947). The use of transformation. *Biometric Bulletin, 3,* 39-52.

Cleveland, W. S. (1984). Graphical methods for data presentation: Full scale breaks, dot charts, and multibased logging. *The American Statistician, 38(4),* 270-280.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Finney, D. J. (1948). Transformation of frequency distributions. *Nature, London, 162,* 898

Goodman, L. A. (1954). Kolmogorov-Smirnov tests for psychological research. *Psychological-Bulletin, 51,* 160-168 Hopkins, W. G. (2002). *A new view of statistics.* Available online at http://www.sportsci.org/resource/stats/index.html.

Lilliefors, H. W. (1968). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association, 62,* 399-402

Judd, C. M., & McClelland, G.H. (1989). *Data analysis: A model-comparison approach.* San Diego, CA: Harcourt Brace Jovanovich.

Orr, J. M., Sackett, P. R., & DuBois, C. L. Z. (1991). Outlier detection and treatment in I/O psychology: A survey of researcher beliefs and an empirical illustration. *Personnel Psychology, 44,* 473- 486.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research.* Harcourt Brace: Orlando, FL.

Rosenthal, R. (1968). An application of the Kolmogorov-Smirnov test for normality with estimated mean and variance. *Psychological-Reports, 22*, 570.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using Multivariate Statistics.* New York: Harper Collins.

Wilcox, R. R. (1997). Some practical reasons for reconsidering the Kolmogorov-Smirnov test. *British Journal of Mathematical and Statistical Psychology, 50(1),* 9-20 Zimmerman, D. W. (1995). Increasing the power of nonparametric tests by detecting and downweighting outliers. *Journal of Experimental Education, 64,* 71-78.

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education, 67,* 55-68.

Zubin, J. (1935). Note on a transformation function for proportions and percentages. *Journal of Applied Psychology, 19,* 213-220.

Author Notes

The author would like to express his gratitude to his former students at the University of Oklahoma for providing him with the impetus to write this paper.

Contact information:

Jason Osborne
ERLCE,

North Carolina State University,
Campus Box 7801
520 Poe Hall
Raleigh, NC 27695-7801

Email: jason_osborne@ncsu.edu

Descriptors: Statistical Distributions; Data Analysis; Nonnormal Distributions; *Parametric Analysis; Skew Curves; Normal Curve

```
ADODB.Recordset error '800a0e78'

Operation is not allowed when the object is closed.

/pare/getvn.asp, line 179
```

Home | Articles | Subscribe | Review | Policies

Volume: 8 7 6 5 4 3 2 1

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ☐ Go

Copyright 2002, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Stansfield, Charles W. (2002). Linguistic simplification: a promising test accommodation for lep students?. *Practical Assessment, Research & Evaluation*, 8(7). Available online: http://ericae.net/pare/getvn.asp?v=8&n=7. This paper has been viewed 2019 times since 7/23/02.

- Find similar papers in
  ERICAE Full Text Library
  Pract Assess, Res & Eval
  ERIC RIE & CIJE 1990-
  ERIC On-Demand Docs
- Find articles in ERIC written by
  Stansfield, Charles W.

Linguistic Simplification: A Promising Test Accommodation for LEP Students?

Charles W. Stansfield
Second Language Testing

In recent years, there has been much discussion about the role of language minority students in state assessments. The vast majority of states surveyed have dealt with the issue by exempting language minority students, with forty-four of forty-eight states exempting limited English proficient (LEP) students from one or more assessments, and more than half (27 of 44) routinely exempting LEP students from state assessments altogether (Rivera et al., 1997). Rivera and Vincent (1997) have questioned the wisdom of this policy. They argue that if LEP students are meant to attain the same high performance standards as their monolingual counterparts, they should be included in state assessments as well. Instead of excluding LEP students from assessments, they argue that states should make judicious use of accommodations that are specially designed with these students' linguistic needs in mind.

There has been little experimental research conducted to investigate the overall effects of accommodations such as those used for students with disabilities, let alone research on accommodations that address the linguistic needs of LEP students. Without empirical data, it is unclear what role a particular test accommodation may play. One accommodation may give an unfair advantage to examinees receiving it, whereas another may not improve the performance of even those who have special needs and should benefit the most from it. Therefore, it is essential that research be conducted to determine whether accommodations are a threat to a test's reliability and validity, or to score comparability for examinees who receive them and examinees who do not.

This article is a synopsis of an experimental study of the effects of linguistic simplification, a test accommodation designed for LEP students. Conducted as part of Delaware's statewide assessment program, this study examined the effects of linguistic simplification of fourth- and sixth-grade science test items and specifically looked at score comparability between LEP and non-LEP examinees.

BEST COPY AVAILABLE

Why Linguistic Simplification? A Review of the Literature on Simplified English

Although the concept of simplifying English has been around for more than seventy years, it has received little attention in research. The first "Basic English" system was designed in 1932 as an alternative, easy means of cross-cultural communication (Ogden, 1932). It consisted of a core vocabulary of 850 words and a few limited syntactic structures.

The concept lay dormant until the 1970s and 1980s, when it was picked up again by multinational corporations looking to facilitate communication and training. Among others, the Caterpillar Corporation (Association Européene de Constructeurs de Matérial Aerospatiale, 1972) and Boeing, Inc. (Shubert et al., 1995) used simplified English to prepare their training manuals for use around the world. Despite its use in corporate settings, only two experimental studies appear to have been conducted on linguistic simplification as an accommodation for LEPs.

Abedi and others (1998) did a study of simplification using mathematics items from the National Assessment of Educational Progress (NAEP). He administered regular NAEP math assessment, a simplified English version, or a Spanish version of the items to 1400 eighth-grade students in southern California middle schools. Results indicated that both LEP and non-LEP students performed best on the simplified version and worst on the Spanish version. However, his analyses also suggested that linguistic simplification doesn't always work as intended, as significant differences in item difficulty were obtained on only 34% of the simplified items. Abedi concluded that linguistic simplification of math items might be beneficial to all students, not just those with limited English proficiency.

Kiplinger et al. (2000) conducted another study using mathematics items from NAEP. This time, a simplified English version, a version with a glossary containing definitions of non-technical terms, and an unsimplified version were administered in Colorado. The instruments were randomly assigned to 1200 special education, LEP, and regular fourth-grade students. Their results showed no significant difference for the three versions across all three types of students, and neither regular nor LEP students performed significantly better on either version. They did find, however, that the students who performed best on the test benefited most from the version that had a glossary, and somewhat from the simplified version. On the basis of these findings, the researchers concluded that glossaries and linguistic simplification might benefit all students.

Experimental Study on the Effects of Linguistic Simplification on a Statewide Science Assessment

Rivera and Stansfield (2001) used Abedi (1998) and Kiplinger et al. (2000) as an impetus for further research on linguistic simplification. Both of these previous studies seemed to provide evidence that linguistic simplification of items might be a useful accommodation for LEPs in formal assessment settings. However, Rivera and Stansfield highlighted the need for a formal experimental study to determine the effect linguistic simplification might have on scores for LEP and non-LEP students. Only once score comparability has been established can an accommodation be rightfully endorsed.

The two researchers conducted a study to examine the effects of linguistic simplification on fourth- and sixth-grade science test items used in the Delaware Student Testing Program. At each grade level, four experimental 10-item testlets were included on the operational forms of the science test. Two of the testlets contained regular field test items that had been linguistically simplified, and the other two contained the same field test items written in regular (unsimplified) English. The testlets were randomly assigned to both LEP and non-LEP students throughout the state.

A total of 11,306 non-LEP students and 109 LEP students took one of the forms of the test. Because the number of LEP students was split among the eight forms, the number of LEP students taking each test form was small, ranging from 6 to 23 students. While the researchers caution that due to the limited sample size, nothing can be generalized about linguistic simplification as an aid to LEP students, the findings for the large non-LEP sample are quite clear.

Results of t-tests performed on mean raw scores, analyses of variance (ANOVAs), and post-hoc pairwise comparisons all indicated that overall, there was no significant difference in scores of non-LEP students who took the simplified version as opposed to the regular (unsimplified) one. This is an important finding because it shows that linguistic simplification can be used without fear of providing an unfair advantage to those who receive it, and thereby affecting the comparability of scores across examinees in this condition. Since linguistic simplification is able to reduce the level of English language proficiency needed to comprehend a test item, it is likely that it can reduce the role of language proficiency in achievement test scores in general.

Limitations and Suggestions for Further Research

Other studies should now address the issue of the usefulness of linguistic simplification for LEP students taking formal and high-stakes assessments. If experimental studies involving large samples of LEP students who are randomly assigned to treatments show that those LEP students who receive simplified items perform statistically and meaningfully better than those who receive the unsimplified version of such items, then the utility of linguistic simplification in meeting the needs of LEP test-takers will be established.

In this study, we chose to simplify items on a statewide science assessment. Therefore, the preliminary results we obtained may not hold for other subject areas, and further research is needed to determine the effects of linguistic simplification in other areas such as math and social studies.

While the small sample size did not allow us to address the effectiveness of linguistic simplification for LEPs, the study's results did show that tests and items can be linguistically simplified without compromising score comparability. However, test developers must exercise caution when carrying out the process of linguistic simplification. The result of the process of linguistic simplification must be to make items accessible to LEPs while not altering the difficulty of the content being tested. And at times, in some items, language and content interact to such an extent that simplification is not possible. However, the results of this study suggest that if test developers and researchers are careful in carrying out linguistic simplification, the resulting assessment could address the linguistic needs of the LEP students without compromising the comparability of the scores obtained on the assessment by taking the standard English version.

References

Abedi, J., Lord, C., & Hofstetter, C. (1998). Impact of Selected Background Variables on Students' NAEP Math Performance. Los Angeles: UCLA Center for the Study of Evaluation/National Center for Research on Evaluation, Standards and Student Testing.

Association Européene de Constructeurs de Matérial Aerospatiale (1972). AECMA Simplified English Document: A Guide for the Preparation of Aircraft Maintenance Procedures in the International Aerospace Maintenance Language. *AIA Issue*, Change 4.

Kiplinger, V.L., Haug, C.A., & Abedi, J. (2000). A Math Assessment Should Assess Math, Not Reading: One State's Approach to the Problem. Paper presented at the 30th National Conference on Large Scale Assessment, Snowbird, UT, June 25-28.

Ogden, C.K. (1932). *Basic English, A General Introduction with Rules and Grammar*. London: Paul Treber & Co.

Rivera, C., & Stansfield, C.W. (2001). The Effects of Linguistic Simplification of Science Test Items on Performance of Limited English Proficient and Monolingual English-Speaking Students. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Rivera, C., Vincent, C., Hafner, A., & LaCelle-Peterson, M. (1997). Statewide Assessment Programs: Policies and Practices for the Inclusion of Limited English Proficient Students. *ERIC Digest.* Washington D.C.: ERIC Clearinghouse on Language and Linguistics. ED 362 073.

Rivera, C., & Vincent, C. (1997). High school graduation testing: Policies and practices in the assessment of English language learners. *Educational Assessment*, 4 (4): 335-55.

Shubert, J.K., et al. (1995). "The Comprehensibility of Simplified English in Procedures." *Technical Writing and Communication,* 25 (4): 347-369.

Descriptors: Second Languages; Student Evaluation; Scoring; Test Construction; Intercultural Communication; Second Language Instruction; Student Motivation

```
ADODB.Recordset error '800a0e78'

Operation is not allowed when the object is closed.

/pare/getvn.asp, line 179
```

Home | Articles | Subscribe | Review | Policies

Volume: 8 7 6 5 4 3 2 1

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal. ISSN 1531-7714

Search: title Go

Copyright 2002, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Mantero, Miguel (2002). Evaluating classroom communication: in support of emergent and authentic frameworks in second language assessment. *Practical Assessment, Research & Evaluation*, 8(8). Available online: http://ericae.net/pare/getvn.asp?v=8&n=8.
This paper has been viewed 3915 times since 7/25/02.

Find similar papers in
ERICAE Full Text Library
Pract Assess, Res & Eval
ERIC RIE & CIJE 1990-
ERIC On-Demand Docs

Find articles in ERIC written by
Mantero, Miguel

Evaluating Classroom Communication: In Support of Emergent and Authentic Frameworks in Second Language Assessment

Miguel Mantero
The University of Alabama

This paper addresses sociocultural theory and pedagogy (Vygotsky 1978, Lantolf 2000) in the second language classroom, particularly as it relates to student assessment. While teaching practices may be evolving to reflect the theory, methods of assessment are still largely the same: based on *a priori* structures and grammar (Hopper and Thompson 1993). Authentic assessment (Wiggins 1990) and instructional conversations (Tharp and Gallimore 1988) are introduced as better methods for student assessment in language classrooms that operate within the sociocultural framework.

Introduction

VanPatten (1998) points out that there is a gap between second language acquisition (SLA) theory and classroom practice due to varying interpretations of the concept 'communicative.' One version of 'communicative' is text driven; communicative activities are provided as the end result of a chapter or segment. Students first 'learn' the material, then they use it to communicate[1]. According to this definition, communicative tasks are a measure of student learning rather than a means by which to acquire language.

An alternative version of 'communicative' is suggested by activity theory (Vygotsky 1978, Wertsch 1991), in which contextualized communicative tasks lead to the acquisition of a language. Indeed, VanPatten's (1998) other definition of 'communicative' ties language acquisition to communicative events in which there is a negotiation of meaning that does not rely on an *a priori* knowledge of grammar. According to this definition, communication is not the result of knowing a grammar; rather, grammar is acquired through communication.

Beliefs about the role and concept of grammar in communication and interaction influence not only SLA theories but also the practices of many foreign language classrooms. Following an overview of sociocultural theory, this article will examine two prevalent views of grammar in foreign language classrooms and explore the assessment implications of each.

Sociocultural Framework

Our minds are mediated by the social, historical, and cultural contexts that surround us at any given moment (Luria 1981). As the main players in the worlds that we inhabit, we change and influence the contexts to suit our understandings and purposes or those of others whom we believe to be valuable. In order to change or influence the worlds in which we live, we use language as the main tool to help us appropriate knowledge or understanding (Volosinov 1973).

According to Vygotsky (1978), Lantolf (2000), and Wertsch (1991), speaking and thinking are not one in the same. Through language, however, we can assess (in everyday contexts) what may be inside someone's mind. Language may be observed as utterance, dialogue, or discourse (Mantero 2002a, 2002b). An utterance, according to Bakhtin (1986), carries with it the possibility of being responded to, and, in turn, creating dialogue. Basically, an utterance is a single spoken "sentence" without a response made by a speaker. It is when the utterance is responded to by another speaker that dialogue is created, and therefore extends beyond the one-sided (monologic). Interaction remains at the dialogue level (i.e., dialogic) if the communication between speakers revolves around one idea. Discourse emerges when dialogue assists in clarifying a new concept. For example, we can exchange utterances (dialogue) about a car in a second language and still remain at the dialogue level of communication. But when we begin to talk about how a car affords us more freedom and how freedom is appreciated by all, then the dialogue surrounding the car has now turned into a discourse on freedom.

True dialogue stems from a negotiation of meaning, an attempt to understand, or convince someone of, a point of view. It does not have a pre-appointed end to it, such as successfully ordering from a menu in an in-class role play. Instead, true dialogue furthers discourse because those involved are using language as a tool in goal-directed action (Tharp and Gallimore 1988, Wells 1999). Within a sociocultural framework, the task of ordering a meal in restaurant would be more effective, discursively, if the students had to negotiate what type of restaurant and food as well as any activities that might happen afterwards. The dialogue of "ordering dinner" would then become embedded within a larger discourse that emerged over time through unrehearsed dialogue that served to activate cognitive processes involved in decision making.

Vygtosky (1978) operationalized dialogue and discourse into areas that he termed the Zone of Actual Development (ZAD) and the Zone of Proximal Development (ZPD). These two concepts are critical for understanding how to approach students in a sociocultural framework.

Basically, the ZAD is what people can do by themselves and the ZPD is what a person can do with assistance from a knowledgeable other. When you go to the corner store for a gallon of milk, you don't need any assistance from others, but when you visit a friend's new house for the first time, you might go with someone who has already been there so that he or she can direct you. Although simplified, this example helps illustrate the differences between the ZAD and the ZPD. It is also important to note that just because a person helps another reach the ZPD once, it doesn't mean that the first person will remember how to get there. In order for the individual ZPD to turn into a ZAD, there must be contextualized, supported practice and action (mental and physical). And here is where the use of tools comes into play. Imagine that you need to go to your friend's new house again, but can't remember how to find the way home. Your friend can't accompany you, so he draws you a map. This map is your tool for working through your ZPD. It replaces the language that you used when you had your friend in the car with you. The more you visit your friend's new house, the less you will need the map or have to ask for directions. Eventually, you'll be able to find the new house without any problems. At this moment, metaphorically at least, your ZAD will have expanded and created a new ZPD (for example, Where can you go from your friend's new house with their help?).

A major implication for language assessment is how we might assess students' use of language if they are mediating through their own continually expanding ZPDs. How will we hold true to the meanings of dialogue and discourse within a language classroom?

Emergent and *a priori* Grammars

Shohamy (2000) states that a reality of foreign language classroom tests is that they identify not what knowing a language means, but what knowing a language means in testing situations. Imagine a typical testing situation in a contemporary classroom. If you envision a classroom filled with students bent over their desks, furiously filling in blanks, listening to a passage then circling the 'right answer,' scribbling verb endings in the margins beside the matching vocabulary section, then you saw the typical assessment methods based on the notion of *a priori* grammar that are prevalent in language curricula. In this situation, knowledge of linguistic structures signals what is understood as a 'communicative student," that is, a monologic student instead of a discursive student. Such a monologic student has not reached beyond his or her ZAD and attempted to expand it. This student stays within the confines that a priori grammar instruction places upon him or her and which is supported, and enforced by, the assessment methods that focus on *a priori* grammar.

If assessment is based on *a priori* grammar knowledge, then the role and process of understanding contexts and culture is diminished and this may lead to a lack of dialogue and discourse within a language classroom. As Saville-Troike (1991) mentions, cultural and contextual knowledge assists students in negotiating meaning, thus entering into dialogue and discourse. Think, for example, of the cultural scripts we use when we go to the bank, order a pizza, or buy a car.

The notion of grammar being acquired discursively through negotiating communicative tasks is consistent with sociocultural theory. Hopper (1993) refers to it as Emergent Grammar. According to this view, grammar is seen as incomplete and in process or emergent. It is not a fixed set of rules one must know in order to do well on a test because the current chapter 'covered' the past tense. Meaning is taken to be contextual. Symbols, linguistic or not, do not require a grammar to be meaningful.

The popular view of grammar in many foreign language classrooms and texts is, however, that which Hopper labels *a priori*. This *a priori* grammar is perfectly monologic, and at the utterance level of classroom interaction within a sociocultural framework (Wells 1999). In order to understand or learn an a priori grammar, we need not involve ourselves in discourse. *a priori* grammar knowledge is easily assessed in classrooms and is used to label students as being less or more 'communicative' than others or 'knowing more' Spanish, French, German, etc.

Littlewood (1980) assists us in viewing the validity of framing our assessment methods through Emergent Grammar by defining linguistic structure as a form that appears through action and interaction. Meaning is not the product of automatic, predisposed blueprints of language. Meaning is contingent on dialogue (Bereiter 1994). Assessment under this view should be dialogic and discursive and allow for an expansion of the ZAD linguistically and cognitively. Further, doing so will allow for more communication and negotiation of meaning that will produce second language acquisition. Two methods for assessing students-instructional conversation and authentic assessment-- consistent with Emergent Grammar and sociocultural theory are described below.

Instructional Conversations

Tharp and Gallimore (1989) define an approach to teaching that, in line with sociocultural thought, may be used to assess students while in the process of thinking and learning. Focusing on the role of the teacher in assisting performance, they further clarify the concept of scaffolding (Wertsch 1991), which encompasses the following seven activities: modeling, providing feedback, applying contingency management (rewards and punishments), directing, questioning, explaining, and structuring tasks. Each of these activities is built within and around the students' ZAD and ZPD, therefore making assessment a part of the process of learning and thinking. Tharp and Gallimore clarify the instructional conversation as such:

> *Parents and teachers who engage in instructional conversations are assuming that the (student) may have something to say beyond the known answers in the head of the adult. They occasionally extract from the (student) a 'correct' answer, but to grasp the communicative intent... adults need to listen carefully, and... to adjust their responses to assist (their) efforts* (p.24).

A student's grammar is dynamically assessed throughout discourse and communication when using any one (or a combination of more than one) of the seven approaches to instruction that Tharp and Gallimore propose to assist the students in discursive, goal-directed action. The instructional conversation, as outlined, provides a framework for this type of assessment and interaction.

Authentic Assessment

To provide an effective method of assessing students' language and cognition, it is helpful to outline a method that has been proposed by Wiggins (1990) and Archbald and Newman (1989): authentic assessment. Authentic assessment is any type of assessment that requires students to demonstrate skills and competencies that realistically represent problems and situations likely to be encountered in daily life. When authentic assessment is placed into the context of a language classroom, what follows is a cognitively more demanding method of assessment that has to include more discourse and reliance on emergent grammar by both the student and the instructor because, as Wiggins states, authentic assessment offers opportunities to plan and revise dialogue and discourse, collaborate with others, and help students 'play' within contextualized worlds inside of the classroom that are based on the culture(s) of the language being studied. Given the very nature of this type of assessment, it complements the sociocultural theory to which many language classrooms are attempting to subscribe.

Implications for Teachers and Teacher Educators

Assessment of language learning can be understood as evaluating either the process of language learning or the product of studying a second language. Instructors need to have a clear vision of what they are assessing: process or product. This may translate into formative or summative assessment in the language classroom.

Traditionally, the American Council on the Teaching of Foreign Languages' Oral Proficiency Interview (OPI) and the Teachers of English to Speakers of Other Language's English as a Second Language (ESL) Standards are tools that are used for summative assessment on *a priori* structures. This limited use places unnecessary constraints on valuable rubrics that the field of language education has relied on and used for a decades. Why evaluate only what a student can do at a given moment of linguistic proficiency with *a priori* constructs? If we keep in mind the goals of authentic assessment and instructional conversations, then rubrics such as the OPI and the ESL Standards can be implemented in a formative manner. Language learning is a process that involves specific feedback from assessment instruments about the student's potential language proficiency as well as actual. When placed in an emergent framework (as discussed earlier), the OPI and ESL Standards assist the students in understanding what they have learned and what they may still need on focus on.

Viewing linguistic proficiency as emergent allows for the assessment methods to be applied in a more formative aspect, and this in turn allows for a truer picture of second language acquisition and learning within the classroom environment. New teachers and teacher-educators will have to decide whether to focus their assessment skills and rubrics to the student's ZAD or ZPD. Assessment based solely in the ZAD focuses on the *a priori* constructs mentioned earlier and often becomes driven by texts and grammar structure. By focusing assessment in the ZPD, an instructor has to take into account the cognitive and linguistic abilities and skills that a student may have, which allows for more self-expression, creation of meaning, and negotiation during communication.

Note

[1] This paper focuses on verbal communication or production rather than written communication and reading because verbal communication requires speakers to negotiate meaning with each other consistent with sociocultural theory.

References

Archbald, D. and Newman, F. (1989). The functions of assessment and the nature of authentic academic achievement, in Berlak (ed.) *Assessing Achievement: Toward the Development of a New Science in of Educational Testing*. Buffalo, NY: SUNY Press.

Bakhtin, M.M. (1986). *Speech Genres and Other Late Essays*. Austin: University of Texas Press.

Bereiter, C. (1994). Implications of postmodernism for science or science as progressive discourse. *Educational Psychologist*, 29 (1): 3-12.

Hopper, P., and Thompson, S. (1993). Language universals, discourse pragmatics, and semantics. *Language Sciences*, 15 (4): 357-76 .

Lantolf, J.P., ed. (2000). *Sociocultural Theory and Second Language Learning* (133-153). New York: Oxford University Press.

Littlewood, W. T. (1980). Form and meaning in language-teaching methodology. *Language Journal*, 64 (4): 441-445.

Luria, A.R. (1981). *Language and Cognition*. Ed. J.V. Wertsch. New York: Wiley Publishing.

Mantero, M. (2002a). *The Reasons We Speak: Cognition and Discourse in the Second Language Classroom*. Westport, CT: Bergin and Garvey.

Mantero, M. (2002b). Bridging the gap: Discourse in text-based foreign language classrooms. *Foreign Language Annals* (in press).

Saville-Troike, M. (1991). Teaching and testing for academic achievement: The role of language development. *Occasional Papers in Bilingual Education*, 4, Spring.

Shohamy, E. (2000). The relationship between language testing and second language acquisition, revisited. *System*, 28 (4): 541-53.

Tharp, R.G., & Gallimore, R. (1989). Rousing Schools to Life. *American Educator*, 13 (2): 20-25, 46-52.

Tharp, R.G., & Gallimore, R. (1988). *Rousing Minds to Life: Teaching, Learning, and Schooling in Social Context*. Cambridge: Cambridge University Press.

VanPatten, B. (1998). Perceptions of and perspectives on the term "communicative." *Hispania*, 81 (4): 925-32.

Volosinov, V.N. (1973). *Marxism and the Philosophy of Language*. New York: Seminar Press.

Vygotsky, L.S (1978). *Mind in Society*. Cambridge, MA: Harvard University Press.

Wells, Gordon. (1999). *Dialogic Inquiry: Toward Sociocultural Practice and Theory of Education*. Cambridge: Cambridge University Press.

Wertsch, J.V. (1991). *Voices of the Mind: A Sociocultural Approach to Mediated Action*. Cambridge, MA: Harvard University Press.

Wiggins, G. (1990). The case for authentic assessment. *Practical Assessment, Research & Evaluation,* 2 (2).

Descriptors: Classroom Communication; Second Languages; Student Evaluation; Scoring; Test Construction; Classroom Techniques; Elementary Secondary Education; Intercultural Communication; Second Language Instruction; Student Motivation

```
ADODB.Recordset error '800a0e78'

Operation is not allowed when the object is closed.

/pare/getvn.asp, line 179
```

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal. ISSN 1531-7714

Copyright 2002, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Boston, Carol (2002). The concept of formative assessment. *Practical Assessment, Research & Evaluation*, 8(9). Available online: http://ericae.net/pare/getvn.asp?v=8&n=9.
This paper has been viewed 11545 times since 8/6/02.

▸ Find similar papers in
ERICAE Full Text Library
Pract Assess, Res & Eval
ERIC RIE & CIJE 1990-
ERIC On-Demand Docs

▸ Find articles in ERIC written by
Boston, Carol

The Concept of Formative Assessment

Carol Boston
ERIC Clearinghouse on Assessment and Evaluation
University of Maryland, College Park

While many educators are highly focused on state tests, it is important to consider that over the course of a year, teachers can build in many opportunities to assess how students are learning and then use this information to make beneficial changes in instruction. This diagnostic use of assessment to provide feedback to teachers and students over the course of instruction is called formative assessment. It stands in contrast to summative assessment, which generally takes place after a period of instruction and requires making a judgment about the learning that has occurred (e.g., by grading or scoring a test or paper). This article addresses the benefits of formative assessment and provides examples and resources to support its implementation.

Purpose and Benefits of Formative Assessment

Black and Wiliam (1998b) define assessment broadly to include all activities that teachers and students undertake to get information that can be used diagnostically to alter teaching and learning. Under this definition, assessment encompasses teacher observation, classroom discussion, and analysis of student work, including homework and tests. Assessments become formative when the information is used to adapt teaching and learning to meet student needs.

When teachers know how students are progressing and where they are having trouble, they can use this information to make necessary instructional adjustments, such as reteaching, trying alternative instructional approaches, or offering more opportunities for practice. These activities can lead to improved student success.

Black and Wiliam (1998a) conducted an extensive research review of 250 journal articles and book chapters winnowed from a much larger pool to determine whether formative assessment raises academic standards in the classroom. They concluded that efforts to strengthen formative assessment produce significant learning gains as measured by comparing the average improvements in the test scores of the students involved in the innovation with the range of scores found for typical groups of students on the same tests. Effect sizes ranged between .4 and .7, with formative assessment apparently helping low-achieving students, including students with learning disabilities, even more than it helped other students (Black and Wiliam, 1998b).

Feedback given as part of formative assessment helps learners become aware of any gaps that exist between their desired goal and their current knowledge, understanding, or skill and guides them through actions necessary to obtain the goal (Ramaprasad, 1983; Sadler, 1989). The most helpful type of feedback on tests and homework provides specific comments about errors and specific suggestions for improvement and encourages students to focus their attention thoughtfully on the task rather than on simply getting the right answer (Bangert-Drowns, Kulick, & Morgan, 1991; Elawar & Corno, 1985). This type of feedback may be particularly helpful to lower achieving students because it emphasizes that students can improve as a result of effort rather than be doomed to low achievement due to some presumed lack of innate ability. Formative assessment helps support the expectation that all children can learn to high levels and counteracts the cycle in which students attribute poor performance to lack of ability and therefore become discouraged and unwilling to invest in further learning (Ames, 1992; Vispoel & Austin, 1995).

While feedback generally originates from a teacher, learners can also play an important role in formative assessment through self-evaluation. Two experimental research studies have shown that students who understand the learning objectives and assessment criteria and have opportunities to reflect on their work show greater improvement than those who do not (Fontana & Fernandes, 1994; Frederikson & White, 1997). Students with learning disabilities who are taught to use self-monitoring strategies related to their understanding of reading and writing tasks also show performance gains (McCurdy & Shapiro, 1992; Sawyer, Graham, & Harris, 1992).

Examples of Formative Assessment

Since the goal of formative assessment is to gain an understanding of what students know (and don't know) in order to make responsive changes in teaching and learning, techniques such as teacher observation and classroom discussion have an important place alongside analysis of tests and homework.

Black and Wiliam (1998b) encourage teachers to use questioning and classroom discussion as an opportunity to increase their students' knowledge and improve understanding. They caution, however, that teachers need to make sure to ask thoughtful, reflective questions rather than simple, factual ones and then give students adequate time to respond. In order to involve everyone, they suggest strategies such as the following:

- Invite students to discuss their thinking about a question or topic in pairs or small groups, then ask a representative to share the thinking with the larger group (sometimes called think-pair-share).
- Present several possible answers to a question, then ask students to vote on them.
- Ask all students to write down an answer, then read a selected few out loud.

Teachers might also assess students' understanding in the following ways:

- Have students write their understanding of vocabulary or concepts before and after instruction.
- Ask students to summarize the main ideas they've taken away from a lecture, discussion, or assigned reading.
- Have students complete a few problems or questions at the end of instruction and check answers.
- Interview students individually or in groups about their thinking as they solve problems.
- Assign brief, in-class writing assignments (e.g., "Why is this person or event representative of this time period in history?)

(The November/December 1997 issue of *Clearinghouse* magazine is devoted to practical ideas for formative assessment. See especially Mullin and Hill for ideas for history classes, McIntosh for mathematics, Childers and Lowry for science, and Bonwell for higher education.)

In addition to these classroom techniques, tests and homework can be used formatively if teachers analyze where students are in their learning and provide specific, focused feedback regarding performance and ways to improve it. Black and Wiliam (1998b) make the following recommendations:

- Frequent short tests are better than infrequent long ones.
- New learning should be tested within about a week of first exposure.
- Be mindful of the quality of test items and work with other teachers and outside sources to collect good ones.

Portfolios, or collections of student work, may also be used formatively if students and teachers annotate the entries and observe growth over time and practice (Duschl & Gitomer, 1997).

Resources for Teachers Interested In Formative Assessment

Formative assessment is tightly linked with instructional practices. Teachers need to consider how their classroom activities, assignments, and tests supports learning aims and allow students to communicate what they know, then use this information to improve teaching and learning. Two practitioner-oriented books that offer many helpful ideas about, and examples of, classroom assessments are *A Practical Guide to Alternative Assessment* (Herman, Aschbacher, and Winters, 1992) and *Classroom Assessment Techniques: A Handbook for College Teachers* (Angelo and Cross, 1993).

The Northwest Regional Educational Laboratory has put large sections of its helpful training kit, *Improving Classroom Assessment: A Toolkit for Professional Developers* online at http://www.nwrel.org/assessment/toolkit98.asp. The readings, overheads, exercises, and handouts could help groups of teachers think through assessment issues in their schools. The Assessment Training Institute provides some free newsletter and journal articles about classroom assessment on its Web site (http://www.assessmentinst.com/) as well as publications, videos, and training sessions for a fee. A recent issue of the *Maryland Classroom* newsletter from the Maryland State Department of Education features a lead article on effective feedback in the classroom with example responses from an assignment involving persuasive text (http://www.msde.state.md.us/Maryland%20Classroom/2002_05.pdf).

The National Research Council (2001) has produced a useful, accessible book on classroom assessment in science that contains many interesting vignettes about how teachers can adjust their teaching based on their observations, questioning, and analysis of student work. While the anecdotes are specific to K-12 science teaching, the chapters about the documented value of formative assessment on classroom achievement, as well as what it requires in terms of teacher development and how classroom assessment relates to summative assessment such as state tests, have broad applicability. See http://www.nap.edu/catalog/9847.html for a browsable version of *Classroom Assessment and the National Science Education Standards.*

Training and professional development in the area of classroom assessment are essential in order to provide individual teachers with the time and support necessary to make changes. Teachers need time to reflect upon their assessment practices and benefit from observing and consulting with other teachers about effective practices and about changes they would like to make (NRC, 2001). Black and Wiliam (1998b) recommend setting up local groups of schools-elementary and secondary; urban, suburban, and rural-to tackle formative assessment at the school level while collaborating with other local schools. They anticipate that challenges will be different in different subject areas and suggest that external evaluators could help teachers with their work and collect evidence of effectiveness. They also point to potential conflicts between state assessments and classroom assessments, where the external tests can shape what goes on in the classroom in a negative way if the emphasis is on drill and test preparation versus teachers' best judgment about learning.

Teachers generally need to undertake or participate in some summative assessment as a basis for reporting grades or meeting accountability standards. However, the task of summative assessment for external purposes remains quite different from the task of formative assessment to monitor and improve progress. While state tests provide a snapshot of a student's performance on a given day under test conditions, formative assessment allows teachers to monitor and guide students' performance over time in multiple problem-solving situations. Future research might examine how teachers deal with the relationship between their formative and summative roles, how teachers' classroom assessments relate to external test results, and how external test results can be made more helpful in terms of improving student performance.

References

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology*, 84 (3): 261-271.

Angelo, T.A., and Cross, K.P. (1993). *Classroom Assessment Techniques: A Handbook for College Teachers*, 2nd ed. San Francisco: Jossey-Bass.

Bangert-Drowns, R.L., Kulick, J.A., and Morgan, M.T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research,* 61 (2): 213-238.

Black, P., and Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education*, 5 (1): 7-74.

Black, P. and Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80 (2): 139-148. (Available online: http://www.pdkintl.org/kappan/kbla9810.htm.)

Bonwell, C.C. (1997). Using active learning as assessment in the postsecondary classroom. *Clearing House*, 71 (2): 73-76.

Childers, P., and Lowery, M. (1997). Engaging students through formative assessment in science. *Clearing House*, 71 (2): 97-102.

Duschl, R.D. and Gitomer, D.H. (1997). Strategies and challenges to change the focus of assessment and instruction in science classrooms. *Educational Assessment*, 4 (1): 37-73.

Elawar, M.C., and Corno, L. (1985). A factorial experiment in teachers' written feedback on student homework: Changing teacher behaviour a little rather than a lot. *Journal of Educational Psychology*, 77 (2): 162-173.

Fontana, D., and Fernandes, M. (1994). Improvements in mathematics performance as a consequence of self-assessment in Portuguese primary school pupils. *British Journal of Educational Psychology*, 64 (3): 407-417.

Frederiksen, J.R., and White, B.J. (1997). Reflective assessment of students' research within an inquiry-based middle school science curriculum. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Herman, J.L., Ashbacher, P.R., and Winters, L. (1992). *A Practical Guide to Alternative Assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

McCurdy, B.L., and Shapiro, E.S. (1992). A comparison of teacher monitoring, peer monitoring, and self-monitoring with curriculum-based measurement in reading among students with learning disabilities. *Journal of Special Education*, 26 (2): 162-180.

McIntosh, M.E. (1997). Formative assessment in mathematics. *Clearing House*, 71 (2): 92-97.

Mullin, J., and Hill, W. (1997). The evaluator as evaluated: The role of formative assessment in history class. *Clearing House,* 71 (2): 88-92.

National Research Council (2001). *Classroom Assessment and the National Science Education Standards*, edited by J.M. Atkin, P. Black, and J. Coffey. Washington, D.C.: National Academy Press. (Browse online at: http://www.nap.edu/catalog/9847.html.)

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science*, 28 (1): 4-13.

Sadler, D.R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18 (2): 119-144.

Sawyer, R. J., Graham, S., and Harris, K.R. (1992). Direct teaching, strategy instruction, and strategy instruction with explicit self-regulation: Effects on the composition skills and self-efficacy of students with learning disabilities. *Journal of Educational Psychology*, 84 (3): 340-352.

Vispoel, W.P., and Austin, J.R. (1995). Success and failure in junior high school: A critical incident approach to understanding students' attributional beliefs. *American Educational Research Journal*, 32 (2): 377-412.

Descriptors: Active Learning; Student Evaluation; Educational Strategies; Learning Strategies;Measures [Individuals]

Practical Assessment, Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Copyright 2003, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Lissitz, Robert W. & Huynh Huynh (2003). Vertical equating for state assessments: issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research & Evaluation*, 8(10). Available online: http://ericae.net/pare/getvn.asp?v=8&n=10. This paper has been viewed 591 times since 4/30/03.

▶ Find similar papers in
ERICAE Full Text Library
Pract Assess, Res & Eval
ERIC RIE & CIJE 1990-
ERIC On-Demand Docs

▶ Find articles in ERIC written by
Lissitz, Robert W.
Huynh Huynh

Vertical Equating for State Assessments: Issues and Solutions in Determination of Adequate Yearly Progress and School Accountability

Robert W. Lissitz, University of Maryland
Huynh Huynh, University of South Carolina

(The authors contributed equally to the paper and their names were listed randomly.)

Of all the provisions of the federal No Child Left Behind (NCLB) legislation, the definition and determination of adequate yearly progress (AYP) is perhaps the most challenging. NCLB requires states to administer reading and mathematics assessments at least once each year to students in grades 3 through 8 (and once more within grades 10-12) by 2005-06, and adds a science assessment administered at least once in each of three grade spans by 2007-08. States may select their own assessments and define their own proficiency levels, but they must submit plans to the U.S. Department of Education to establish goals for what percentages of students in various subgroups (e.g., low income, minority, limited English proficient) will meet or exceed proficiency levels on the state's assessments each year.

This article describes AYP and some of the psychometric issues it raises. It examines scaling as a means to equate tests as part of a process to confirm educational gains. Vertically moderated standards are recommended over vertical equating of state assessments to measure annual progress and provide useful instructional information.

ADEQUATE YEARLY PROGRESS

In a paper titled "Making Valid and Reliable Decisions in Determining Adequate Yearly Progress" (Marion et al., 2002), the Council of Chief State School Officers summarizes AYP as follows:

> *Each of at least 9 subgroups of students must reach proficient or advanced achievement levels in reading or language arts and mathematics by 2013-2014 (Uniform progress is required beginning in 2002-03.) AYP determinations are based solely on student achievement results on State assessments. At least 95% of the students in each subgroup must participate in the assessments and all must meet the State's performance target in another academic indicator as prescribed by the law* (p. 5).

Further,

> *The NCLB Act requires States to determine the number of students in a group necessary to yield statistically reliable information as well as the number of students required to be in a group to ensure that the results will not reveal personally identifiable information about an individual student* (p. 12).

BEST COPY AVAILABLE

To briefly summarize the challenge, NCLB requires states to develop a system that tracks students' (by defined subgroups) success in reading/language arts and mathematics (with science coming on board soon) as the students progress through school, and the data associated with these adequately sized subgroups must show at least minimum levels of proficiency. Some additional indicators must be provided, as well, but the primary focus will be on the determination of proficiency and the success of most students over the time span of schooling. The purpose of the AYP is to allow the state to monitor progress and to identify problem schools, and low performing subgroups and to prescribe remediation that will result in No Child Left Behind.

PSYCHOMETRIC ISSUES RELATED TO AYP DETERMINATION

Since test scores are going to be the major source for determining student progress and school accountability under NCLB, it is critical that test scores be comparable from test to test and year to year. Scaling is a measurement technique that can facilitate test score comparability.

Description of Scaling

A scaling process, in general terms, is one in which raw scores (usually calculated as the total number of correct responses) are transformed to a new set of numbers with certain selected attributes, such as a particular mean and standard deviation. For example, the Scholastic Aptitude Test has scores that range from 200 to 800 and result from a scaling process that transforms the number correct score that a student has obtained. Some scaling procedures are non-linear transformations of the raw scores and some are linear. The particular approach used depends upon the purpose of the scaling and the properties that we want in the resulting scale.

One of the most common purposes of scaling has to do with equating two or more tests. The tests to be equated might be given at different times, so that the purpose of the scaling would be to arrive at comparable scores for tests across time. The tests might be given to different groups, as well. The most common application for scaling involves equating different forms of the same test. In any case, the rescaling of the students' raw score performance level has the following advantages:

- Regardless of changes in the test from year to year, the scores reported to the public are always on the same scale. This makes it easier for teachers and principals, as well as students and parents, to learn to interpret the results of testing.

- If several related tests need to be available for use, transforming each one to the same scale allows them all to be interpreted in a similar way. Again, this helps the problem of communication of test results.

- Equal raw scores from different forms will not usually express the same amount of ability because one form might be easy but the other form might be more difficult. Scaling allows us to "equate" the two forms for purposes of reporting.

Two primary situations exist for scaling multiple sets of tests to a common scale or equating them. Horizontal equating is designed to test different groups of students that are assumed to be at approximately the same level. It occurs within grade, where multiple forms are used to test the same general content. Vertical equating may be used when testing students who are at different levels of education. It entails across-grade testing of the same general content. Each type is discussed in more detail below.

Within-Grade (Horizontal) Scaling. An example of the horizontal equating situation is the case in which a school system has a test for graduation and students are allowed to retake the test if they fail. The retakes are on different forms of the test that are all equated to provide comparable scores. The cut-off for failing is set at the same scale score level no matter how often a student retakes the test, thus ensuring a constancy of the standard for passing from administration to administration. The table of specifications (i.e., the test blueprint) for each test is also the same, thus ensuring that content is comparable and that the dimensions of knowledge that underlie the test are the same in each case. The difficulty level will be approximately the same for each form of the test, as well. Occasionally, horizontal equating is used to allow for comparison of groups of students that are different in some fundamental way that requires modification of one form of the test. For example, comparisons of recent immigrant students who speak only Spanish with those who are fluent in English will require tests that are equated, yet differ in the language of the test items.

Across-Grade (Vertical) Scaling. One of the common ways that psychometricians have approached the AYP problem is to develop a single (unidimensional) scale that summarizes the achievement of students. This scale is then used to directly compare the performance level across grade levels. For example, TerraNova K-12 (CTB/McGraw-Hill, 1997, 2001), the Stanford Achievement Test from Harcourt (1996), and the recent work in Mississippi (Tomkowicz and Schaeffer, 2002) present scales that are purported to allow for the meaningful, continuous, tracking of students across grades.

A classic example of the vertical equating situation is that of a test of mathematics that is used to track expertise across middle school. In this scenario, the tests at different grade levels are of differing content, but still focus on the same general concept, say, mathematics fluency. The students are expected to show performance improvements at each year, and these improvements should be reflected in a steady increase in their ability to do mathematics. The tests for grades 7 and 8 should be linked so that scores are directly comparable along a common continuum or dimension. Sometimes this approach is used for tests of literacy, as well.

The content must have some sense of commonality across grades in order to be meaningfully equated across grade levels. These

BEST COPY AVAILABLE

scales are often considered developmental, in the sense that they encourage the examination of changes in a student's score across grades that indicate the improvement in that student's competency level. Sometimes the equating is only for adjacent grades and sometimes equating is across the whole school experience.

Major Assumptions for Horizontal and Vertical Scaling

The major assumption for equating is that the tests are assessing the same general content. In other words, a psychometric model will be appropriate for each test being scaled or equated and it will develop the modeling relating the two tests using a single or a common set of dimensions. In the case of horizontal scaling, this is not usually a problem. Since each form of the test is designed to examine the same curriculum material, a model that works for one test will usually work for all the forms of that test. Naturally, we are assuming that the tests are not only designed with the same table of specifications, but are using the same mix of test item types. For example, each test would have approximately the same mixture of performance items and selected response items. The English language demands would be about the same, as well. In situations such as these, we have had considerable success with modeling and achieving quite accurate equating (i.e., successful scaling).

Vertical scaling has the same assumption of comparable content. It is assumed that the same basic dimension or dimensions are being assessed in each grade for which we are developing a test to be equated to other grades. This implies that the same dimensions are the focus of the teacher's efforts in each grade, as well. This is usually a problem if the goal is to scale across more than two adjacent grades. Even with two adjacent grades, it is not usually clear that the same dimensions are being assessed. If you are trying to scale two or more tests and the tests are really not assessing the same content, you are actually predicting one from the other, rather than equating the two.

The equating of two tests in the horizontal scaling context is fairly easy using an item response theory (IRT) approach (e.g., Stocking and Lord, 1983). If one believes that the content dimensionality assumption in vertical equating is met, then a variety of approaches can be adopted to accomplish the task of equating across several grades. The paper by Tomkowicz and Schaeffer (2002) about implementing a strategy in Mississippi provides one example. Vertical equating also has been carried out on a trial basis for the South Carolina PACT assessments in reading and mathematics. Generally, the procedures focus on adjacent grades since these are usually the most instructionally similar and more likely to be content similar, as well. The successful equating across grades also involves careful design of each grade's test so that overlap across grades will be more systematically achieved. For example, to vertically equate grades 3 through 8, the design for each test will involve carefully crafted subtests (one for grades 3 and 8 and two for the other grades). This will provide enough overlap in difficulty level to allow scaling adjacent grades.

Major Problems with Vertical Equating

Vertical equating is useful mainly in reading and mathematics, the two subjects that are taught and learned continuously through the schooling process. A vertically equated scale cannot be reasonably constructed for subjects like science (e.g., trying to equate physics and geology) or social studies, and issues arise even in scaling mathematics or reading/language arts. A vertical scale captures the common dimension(s) across the grades; it does not capture grade-specific dimensions that may be of considerable importance. The instructional expectations for teaching and learning reading/language arts and mathematics may not really be summarized by one (or even a few) common dimensions across grades.

The assumption of equal-interval measurements within a grade is not easily met either, and across grades it is very hard to justify, so the comparison of growth at different times in a student's life or comparisons of different groups of students at different grades cannot be satisfactory made. Since the typical motivation for vertically equated scales revolves around capturing the developmental process, this difficulty is a serious issue for schools wishing to implement vertical equating.

Going to a single dimension to capture a very rich assessment environment encourages simplifications that lose the very insights that the assessments were done to illuminate. As Haertel (1991) noted with regard to the decision to abandon across-grade scaling for grades 4, 8, and 12 for the National Assessment of Educational Progress, "In fact, it is very difficult to say anything useful about the fact that eighth graders outperform fourth graders by more points than twelfth graders outperform eighth graders" (p. 13).

Since the nature of the items and the assessment process often changes over grades, vertical equating mixes or confounds content changes with method changes. This makes interpretation of results difficult and violates the assumption of comparable assessment across grades. Further, capturing the span of test difficulty within a single scale is very difficult.

Creating the vertical scale is also a technically difficult task, even with (perhaps because of) the use of IRT models. Artificial adjustments must be made to smooth out the results. As Camilli (1999) indicates, "Dimensionality remains a concern, though investigation of its interaction with equating is significantly complicated by indeterminacy of the latent scale" (p. 77). In simplest terms, performance and learning are essentially multidimensional activities.

AN ALTERNATIVE APPROACH: VERTICALLY MODERATED STANDARDS

BEST COPY AVAILABLE

After examining the problems related to vertical scaling, it is reasonable to conclude that the construction of a vertical scale to equate state assessments is difficult to accomplish, difficult to justify, and difficult to utilize productively. Even if a satisfactory vertical scale could be constructed, it makes little sense to report reading and mathematics on one vertical scale and science and social studies on a different scale. Within-grade scales could be useful in themselves; however, there is another approach that could be even more beneficial to help teachers and principals use state assessment results as they try to comply with the NCLB legislation.

We recommend vertically moderated standards—a primary focus upon the categories of performance that a given state department of education has determined (e.g., advanced, proficient, basic, below basic) and efforts to relate these explicitly to adequate yearly progress through a carefully crafted judgment process. In other words, we recommend defining AYP in terms of adequate end-of-year performance that enables a student to successfully meet the challenges in the next grade. Vertically moderated standards call for state departments of education to implement a judgmental process and a statistical process that, when coupled, will enable each school to project these categories of student performance forward to predict whether each student is likely to attain the minimum, or proficient, standard for graduation, consistent with NCLB requirements.

With the focus of assessment necessarily upon classroom instruction and teachers' adaptation to student needs, changes in the specific scale scores should not be the focus. Rather, the focus should be upon each student meeting the achievement categories at a level that predicts adequate (i.e., successful) achievement in the next grade. Particularly in a large state assessment, it is important to use a common reporting system for all students and this approach will accomplish that.

General Considerations for Vertically Moderated Standards

Mislevy (1992) and Linn and Baker (1993) defined four types of linking: equating, calibration, projection, and moderation. These are listed in decreasing order in terms of the assumptions required, with equating requiring the strongest assumptions and moderation the weakest. The ordering of the four types is also in decreasing order in terms of the strength of the link produced.

Under the best conditions, vertical scaling would fall under the category of "calibration." Given misgivings about the feasibility and usefulness of vertical scaling for state assessments, we believe that a procedure that combined the major features of "projection" and "moderation" should be considered and a reporting system that emphasizes achievement levels (e.g., similar to the NAEP categories of advanced, proficient, basic, below basic) would provide information that is easier to understand. This may necessitate that states undertake a new round of standard setting for their assessments. We recommend that cut scores for each test be set for all grades such that (a) each achievement level has the same (generic) meaning across all grades, and (b) the proportion of students in each achievement level follow a growth curve trend across these grades.

The first criterion may be referred to as "policy equating" in the sense that a common meaning is attached to each achievement category. Thus, in some sense, the term "equating" is used in the context of a qualitative (i.e., having to do with quality) interpretation of test score. The second criterion is similar to the "linear statistical adjustment" (Mislevy, 1992) that imposes some level of consistency in the normative data of all grades under consideration. This type of consistency is based on the belief that current instructional efforts and expectations are approximately equivalent in all grade levels, so there should not be wild and unpredictable variations in student performance across grades for an entire state.

An Example of Vertically Moderated Standards

The 1999 standard setting for the South Carolina 1999 PACT assessments (Huynh, Meyer, & Barton, 2000) produced standards that may be described as "vertically moderated." The South Carolina process followed three basic steps:

- A common set of policy definitions for the achievement levels was agreed upon for all grades in each area.
- Cut scores were initially set for grades 3 and 8 only.
- Once the final cut scores for these grades were adopted by the state based upon a technical advisory committee's recommendation, cut scores for grades 4 through 7 were interpolated from those of grades 3 and 8. A simple growth curve trend line was used in the interpolation.

Procedures for Developing Vertically Moderated Standards

Setting vertically moderated standards for several grades requires adopting a forward-looking orientation regarding proficiency, examining curriculum across grades, considering smoothing procedures for the statistical process, and paying special attention to issues related to at-risk students. States should also conduct annual validation studies to guide their assessment programs. These procedures are introduced below.

Forward-looking Definition of Proficient. The complexity of this judgment process would indicate that most states would requires two groups—one for mathematics and one for reading/language arts--to provide advice that determines the cut-points that would be used to define levels of achievement on a test. The definition of proficient should be forward-looking; that is, students who achieve that category should be understood to be proficient on the material from the grade covered by that year's end-of-grade testing and also judged to have made adequate yearly progress at a level that will enable them to likely be successful in the context of the next school grade. In other words, students who score at the proficient level on an assessment should have the educational background from that grade to succeed in the next.

BEST COPY AVAILABLE

Use of Content Scatter Plots. The judgment process for determining cut-points for the categories of performance on a state's end-of-year exam will involve examining certain relevant data, in addition to the test items from the end-of-year test. In the new process, the judges will need to see the test that will be used in the next grade's end-of-year exam, as well as a description of the relevant curriculum for both years.

The new process will also require that the judges become familiar with a grade-to-grade scatter plot of the two test blueprints, a so-called assessment scatter plot. The scatter plot presentation will be a comparison of the assessment design from the current grade to the coming grade. This curriculum/test assessment blueprint scatter plot will provide an indication of the topic areas that are found on both exams (for example, mathematics at grade 7 and mathematics at grade 8) as well as the topic areas that are unique to each exam (i.e., that material which has no overlap across grades). Taking into consideration the content scatter plot will help maintain the common qualitative interpretation of the achievement levels across grades.

Use of Smoothing Procedures for Interpolation and/or Extrapolation. To set vertically moderated standards for several grades (say 3 through 8), there may be no need to conduct the standard setting for all grades. This may be done for two grades at a minimum, but perhaps three grades will be necessary. Interpolation and/or extrapolation would then be used to compute the cut scores for the other grades, with an eye on the proportion of students who are judged to be proficient at each grade. We recommend that cut scores be smoothed out so that the proportion of student in each achievement level is reasonably consistent from one grade to the next. A smoothing procedure may prove satisfactory for the statistical process, which would then supplement the professional judgment involved.

Use of Margin of Error with Focus on At-Risk Students. For many large-scale assessment programs (such as NAEP and the South Carolina state assessment), deliberations regarding the final set of cut scores often take into account the margin of error inherent in any standard-setting process. Judges vary in backgrounds and their individual, recommended, cut scores often vary as well. Therefore it is safe to presume that, over a large pool of judges, the (true) recommended cut score would fall within a reasonably small band, centered at the recommended cut score.

For an assessment program with heavy focus on instructional improvement, some attention may need to be paid to the students who are at risk of being in a false positive category. These are students deemed marginally proficient in the current year, but who may not have acquired the necessary skills needed for learning the material that will be presented next year. They may be at risk of not reaching the proficient level, as required by AYP, at the end of the following year. Supplemental data, such as grades, attendance, special education or limited-English-proficient status, and teacher documentation will aid in the identification, and subsequent remediation, of at-risk students.

Annual Validation Study. Each year, state department of education should also do a validation study in order to identify any problems with the implementation or operationalization of the system into school practice and to see if changes in the level of proficiency are warranted to lead to the overall success of the schools at the end of the 2013-2014 year, as mandated by the NCLB. State department of education will also need to identify as early as possible those schools that do not seem to be on track to meet the federal guidelines for success (100% of the students achieving proficiency within 10 years). Appropriate assistance, sanctions, and rewards can then be offered.

Vertically moderated standards show great promise for state departments of education attempting to track student performance and academic growth on state assessments in a way that is responsive to the NCLB requirements and also yields genuinely useful instructional information. The combination of judgment and statistical analysis described in this article should result in the creation of cut scores that describe proficiency both in terms of a student's mastery of grade-level material and the likelihood that s/he will be ready for the academic challenges of the next grade. Where students score below proficient, appropriate remediation can be offered early so that schools meet annual yearly progress goals and, more important, children are not left behind.

Note:

This paper is based on a report originally prepared for the Technical Advisory Committee of the Arkansas Department of Education.

References

Camilli, G. (1999). Measurement error, multidimensionality, and scale shrinkage: A reply to Yen and Burket. *Journal of Educational Measurement,* 36, 73-78.

CTB/McGraw-Hill, (1997, 2001) TerraNova, Monterey, CA: Authors.

Haertel, E. (1991). Report on TRP analyses of issues concerning within-age versus cross-age scales for the National Assessment of Educational Progress. (ERIC Clearinghouse Document Reproduction Service No ED404367): Washington, DC: National Center for Education Statistics.

BEST COPY AVAILABLE

Huynh, H., Meyer, P., & Barton, K. (2000). Technical Documentation for the South Carolina 1999 Palmetto Achievement Challenge Tests of English Language Arts and Mathematics, Grades Three Through Eight . Columbia, SC: South Carolina Department of Education.

Linn, R. L., & Baker, E. L. (1993; Winter). Comparing results from disparate assessments. *The CRESS Line*, pp 1-2. Los Angeles: National Center for Research on Evaluation, Standards, & Student Testing.

Marion, S, et. al. (2002). Making valid and reliable decisions in determining adequate yearly progress. Washington, D.C.: Council of Chief State School Officers.

Mislevy. R. J. (1992). Linking educational assessments: Concepts, issues, methods, and prospects. Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-210.

Harcourt Educational Measurement (1996). Stanford Achievement Test Test Series, Ninth Edition. San Antonio, TX: Authors

Tomkowicz, J., & Schaeffer, G. (2002, April). Vertical scaling for custom criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Descriptors: Accountability; Elementary Secondary Education; Federal Government; Federal Legislation; AYP; Methods; Equating

```
ADODB.Recordset error '800a0e78'

Operation is not allowed when the object is closed.

/pare/getvn.asp, line 179
```

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

☒ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").

EFF-089 (1/2003)